ИЗДАТЕЛЬСТВО
МИР

Б. П. Демидович

И. А. Марон

ОСНОВЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

Издательство

«Наука»

COMPUTATIONAL MATHEMATICS

B. P. DEMIDOVICH, I. A. MARON

TRANSLATED FROM THE RUSSIAN
BY GEORGE YANKOVSKY

MIR PUBLISHERS · MOSCOW

First published 1973
Second printing 1976
Third printing 1981

На английском языке

© English translation, Mir Publishers, 1981

PREFACE

The rapid development of computing machines and the broadening application of modern mathematical techniques to engineering investigations have greatly enhanced demands concerning the mathematical training of engineers and scientific workers who deal with applied problems.

The mathematical education of the investigating engineer can no longer be confined to the traditional departments of the so-called "classical analysis" which was established in its basic outlines at the beginning of this century. Research engineers have to know many areas of modern mathematics and, primarily, require a firm grasp of the methods and techniques of computational mathematics insofar as the solution of almost every engineering problem must be carried to a numerical result.

Present-day computing devices have greatly extended the realm of computational work, making it possible, in many instances, to reject approximate interpretations of applied problems and pass on to the solution of precisely stated problems. This involves the utilization of deeper specialized divisions of mathematics (nonlinear differential equations, functional analysis, probabilistic methods, etc.).

Proper utilization of modern computers is impossible without the skilled use of methods of approximate and numerical analysis. All this explains the universal enhanced interest in the methods of computational mathematics.

The basic aim of this book is to give as far as possible a systematic and modern presentation of the most important methods and techniques of computational mathematics on the basis of the general course of higher mathematics taught in higher technical schools. The book has been arranged so that the basic portion constitutes a manual for the first cycle of studies in approximate computations for higher technical colleges. The text contains supplementary material which goes beyond the scope of the ordinary college course, but the reader can select those sections which interest him and omit any extra material without loss of continuity. The chapters and sections which may be dropped out in a first reading are marked with an asterisk.

This text makes wide use of matrix calculus. The concepts of a vector, matrix, inverse matrix, eigenvalue and eigenvector of a matrix, etc. are workaday tools. The use of matrices offers a number of advantages in presenting the subject matter since they greatly facilitate an understanding of the development of many computations. In this sense a particular gain is achieved in the proofs of the convergence theorems of various numerical processes. Also, modern high-speed computers are nicely adapted to the performance of the basic matrix operations.

For a full comprehension of the contents of this book, the reader should have a background of linear algebra and the theory of linear vector spaces. With the aim of making the text as self-contained as possible, the authors have included all the necessary starting material in these subjects. The appropriate chapters are completely independent of the basic text and can be omitted by readers who have already studied these sections.

A few words about the contents of the book. In the main it is devoted to the following problems: operations involving approximate numbers, computation of functions by means of series and iterative processes, approximate and numerical solution of algebraic and transcendental equations, computational methods of linear algebra, interpolation of functions, numerical differentiation and integration of functions, and the Monte Carlo method.

A great deal of attention is devoted to methods of error estimation. Nearly all processes are provided with proofs of convergence theorems, and the presentation is such that the proofs may be omitted if one wishes to confine himself to the technical aspects of the matter. In certain cases, in order to pictorialize and lighten the presentation, the computational techniques are given as simple recipes.

The basic methods are carried to numerical applications that include computational schemes and numerical examples with detailed steps of solution. To facilitate understanding the essence of the matter at hand, most of the problems are stated in simple form and are of an illustrative nature. References are given at the end of each chapter and the complete list (in alphabetical order) is given at the end of the book.

The present text offers selected methods in computational mathematics and does not include material that involves empirical formulas, quadratic approximation of functions, approximate solutions of differential equations, etc. Likewise, the book does not include material on programming and the technical aspects of solving mathematical problems on computers. The interested reader must consult the special literature on these subjects.

B. P. Demidovich,
I. A. Maron

CONTENTS

PREFACE 5

INTRODUCTION. GENERAL RULES OF COMPUTATIONAL WORK 15

CHAPTER 1

APPROXIMATE NUMBERS 19

1.1 Absolute and relative errors 19
1.2 Basic sources of errors 22
1.3 Scientific notation. Significant digits. The number of correct digits 23
1.4 Rounding of numbers 26
1.5 Relationship between the relative error of an approximate number and the number of correct digits 27
1.6 Tables for determining the limiting relative error from the number of correct digits and vice versa 30
1.7 The error of a sum 33
1.8 The error of a difference 35
1.9 The error of a product 37
1.10 The number of correct digits in a product 39
1.11 The error of a quotient 40
1.12 The number of correct digits in a quotient 41
1.13 The relative error of a power 41
1.14 The relative error of a root 41
1.15 Computations in which errors are not taken into exact account 42
1.16 General formula for errors 42
1.17 The inverse problem of the theory of errors 44
1.18 Accuracy in the determination of arguments from a tabulated function 48
1.19 The method of bounds 50
*1.20 The notion of a probability error estimate 52
References for Chapter 1 54

CHAPTER 2

SOME FACTS FROM THE THEORY OF CONTINUED FRACTIONS 55

2.1 The definition of a continued fraction 55
2.2 Converting a continued fraction to a simple fraction and vice versa 56

2.3 Convergents 58
2.4 Nonterminating continued fractions 66
2.5 Expanding functions into continued fractions 72
References for Chapter 2 76

CHAPTER 3

COMPUTING THE VALUES OF FUNCTIONS 77

3.1 Computing the values of a polynomial. Horner's scheme 77
3.2 The generalized Horner scheme 80
3.3 Computing the values of rational fractions 82
3.4 Approximating the sums of numerical series 83
3.5 Computing the values of an analytic function 89
3.6 Computing the values of exponential functions 91
3.7 Computing the values of a logarithmic function 95
3.8 Computing the values of trigonometric functions 98
3.9 Computing the values of hyperbolic functions 101
3.10 Using the method of iteration for approximating the values of a function 103
3.11 Computing reciprocals 104
3.12 Computing square roots 107
3.13 Computing the reciprocal of a square root 111
3.14 Computing cube roots 112
References for Chapter 3 114

CHAPTER 4

APPROXIMATE SOLUTIONS OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS 115

4.1 Isolation of roots 115
4.2 Graphical solution of equations 119
4.3 The halving method 121
4.4 The method of proportional parts (method of chords) 122
4.5 Newton's method (method of tangents) 127
4.6 Modified Newton method 135
4.7 Combination method 136
4.8 The method of iteration 138
4.9 The method of iteration for a system of two equations 152
4.10 Newton's method for a system of two equations 156
4.11 Newton's method for the case of complex roots 157
References for Chapter 4 161

CHAPTER 5

SPECIAL TECHNIQUES FOR APPROXIMATE SOLUTION OF ALGEBRAIC EQUATIONS 162

5.1 General properties of algebraic equations 162
5.2 The bounds of real roots of algebraic equations 167

5.3 The method of alternating sums 169
5.4 Newton's method. 171
5.5 The number of real roots of a polynomial 173
5.6 The theorem of Budan-Fourier 175
5.7 The underlying principle of the method of Lobachevsky-Graeffe . . 179
5.8 The root-squaring process 182
5.9 The Lobachevsky-Graeffe method for the case of real and distinct roots 184
5.10 The Lobachevsky-Graeffe method for the case of complex roots . . 187
5.11 The case of a pair of complex roots 190
5.12 The case of two pairs of complex roots 194
5.13 Bernoulli's method 198
References for Chapter 5 202

CHAPTER 6

ACCELERATING THE CONVERGENCE OF SERIES 203

6.1 Accelerating the convergence of numerical series 203
6.2 Accelerating the convergence of power series by the Euler-Abel method 209
6.3 Estimates of Fourier coefficients 213
6.4 Accelerating the convergence of Fourier trigonometric series by the method of A. N. Krylov 217
6.5 Trigonometric approximation 225
References for Chapter 6 228

CHAPTER 7

MATRIX ALGEBRA 229

7.1 Basic definitions 229
7.2 Operations involving matrices 230
7.3 The transpose of a matrix 234
7.4 The inverse matrix 236
7.5 Powers of a matrix 240
7.6 Rational functions of a matrix 241
7.7 The absolute value and norm of a matrix 242
7.8 The rank of a matrix 248
7.9 The limit of a matrix 249
7.10 Series of matrices 251
7.11 Partitioned matrices 256
7.12 Matrix inversion by partitioning 260
7.13 Triangular matrices 265
7.14 Elementary transformations of matrices 268
7.15 Computation of determinants 269
References for Chapter 7 272

CHAPTER 8

SOLVING SYSTEMS OF LINEAR EQUATIONS 273

8.1 A general description of methods of solving systems of linear equations 273
8.2 Solution by inversion of matrices. Cramer's rule 273
8.3 The Gaussian method 277
8.4 Improving roots 284
8.5 The method of principal elements 287
8.6 Use of the Gaussian method in computing determinants 288
8.7 Inversion of matrices by the Gaussian method 290
8.8 Square-root method 293
8.9 The scheme of Khaletsky 296
8.10 The method of iteration 300
8.11 Reducing a linear system to a form convenient for iteration . . . 307
8.12 The Seidel method 309
8.13 The case of a normal system 311
8.14 The method of relaxation 313
8.15 Correcting elements of an approximate inverse matrix 316
References for Chapter 8 321

*CHAPTER 9

THE CONVERGENCE OF ITERATION PROCESSES FOR SYSTEMS OF LINEAR EQUATIONS 322

9.1 Sufficient conditions for the convergence of the iteration process 322
9.2 An estimate of the error of approximations in the iteration process 324
9.3 First sufficient condition for convergence of the Seidel process . . 327
9.4 Estimating the error of approximations in the Seidel process by the m-norm 330
9.5 Second sufficient condition for convergence of the Seidel process . . 330
9.6 Estimating the error of approximations in the Seidel process by the l-norm 332
9.7 Third sufficient condition for convergence of the Seidel process . . 333
References for Chapter 9 335

CHAPTER 10

ESSENTIALS OF THE THEORY OF LINEAR VECTOR SPACES 336

10.1 The concept of a linear vector space 336
10.2 The linear dependence of vectors 337
10.3 The scalar product of vectors 343
10.4 Orthogonal systems of vectors 345
10.5 Transformations of the coordinates of a vector under changes in the basis 348
10.6 Orthogonal matrices 350
10.7 Orthogonalization of matrices 351

10.8 Applying orthogonalization methods to the solution of systems of linear equations 358
10.9 The solution space of a homogeneous system 364
10.10 Linear transformations of variables 367
10.11 Inverse transformation 373
10.12 Eigenvectors and eigenvalues of a matrix 375
10.13 Similar matrices 380
10.14 Bilinear form of a matrix 384
10.15 Properties of symmetric matrices 384
*10.16 Properties of matrices with real elements 389
References for Chapter 10 393

*CHAPTER 11

ADDITIONAL FACTS ABOUT THE CONVERGENCE OF ITERATION PROCESSES FOR SYSTEMS OF LINEAR EQUATIONS 394

11.1 The convergence of matrix power series 394
11.2 The Cayley-Hamilton theorem 397
11.3 Necessary and sufficient conditions for the convergence of the process of iteration for a system of linear equations 398
11.4 Necessary and sufficient conditions for the convergence of the Seidel process for a system of linear equations 400
11.5 Convergence of the Seidel process for a normal system 403
11.6 Methods for effectively checking the conditions of convergence . . 405
References for Chapter 11 409

CHAPTER 12

FINDING THE EIGENVALUES AND EIGENVECTORS OF A MATRIX . . . 410

12.1 Introductory remarks 410
12.2 Expansion of secular determinants 410
12.3 The method of Danilevsky 412
12.4 Exceptional cases in the Danilevsky method 418
12.5 Computation of eigenvectors by the Danilevsky method 420
12.6 The method of Krylov 421
12.7 Computation of eigenvectors by the Krylov method 424
12.8 Leverrier's method 426
12.9 On the method of undetermined coefficients 428
12.10 A comparison of different methods of expanding a secular determinant 429
12.11 Finding the numerically largest eigenvalue of a matrix and the corresponding eigenvector 430
12.12 The method of scalar products for finding the first eigenvalue of a real matrix 436
12.13 Finding the second eigenvalue of a matrix and the second eigenvector 439
12.14 The method of exhaustion 443

12.15 Finding the eigenvalues and eigenvectors of a positive definite symmetric matrix 445
12.16 Using the coefficients of the characteristic polynomial of a matrix for matrix inversion 450
12.17 The method of Lyusternik for accelerating the convergence of the iteration process in the solution of a system of linear equations 453
References for Chapter 12 458

CHAPTER 13

APPROXIMATE SOLUTION OF SYSTEMS OF NONLINEAR EQUATIONS 459

13.1 Newton's method 459
13.2 General remarks on the convergence of the Newton process . . . 465
*13.3 The existence of roots of a system and the convergence of the Newton process 469
*13.4 The rapidity of convergence of the Newton process 474
*13.5 Uniqueness of solution 475
*13.6 Stability of convergence of the Newton process under variations of the initial approximation 478
13.7 The modified Newton method 481
13.8 The method of iteration 484
*13.9 The notion of a contraction mapping 487
*13.10 First sufficient condition for the convergence of the process of iteration 491
*13.11 Second sufficient condition for the convergence of the process of iteration 493
13.12 The method of steepest descent (gradient method) 496
13.13 The method of steepest descent for the case of a system of linear equations 501
*13.14 The method of power series 504
References for Chapter 13 506

CHAPTER 14

THE INTERPOLATION OF FUNCTIONS 507

14.1 Finite differences of various orders 507
14.2 Difference table 510
14.3 Generalized power 517
14.4 Statement of the problem of interpolation 518
14.5 Newton's first interpolation formula 519
14.6 Newton's second interpolation formula 526
14.7 Table of central differences 530
14.8 Gaussian interpolation formulas 531
14.9 Stirling's interpolation formula 533
14.10 Bessel's interpolation formula 534
14.11 General description of interpolation formulas with constant interval 536
14.12 Lagrange's interpolation formula 539
*14.13 Computing Lagrangian coefficients 543

14.14 Error estimate of Lagrange's interpolation formula 547
14.15 Error estimates of Newton's interpolation formulas 550
14.16 Error estimates of the central interpolation formulas 552
14.17 On the best choice of interpolation points 553
14.18 Divided differences 554
14.19 Newton's interpolation formula for unequally spaced values of the argument 556
14.20 Inverse interpolation for the case of equally spaced points . . . 559
14.21 Inverse interpolation for the case of unequally spaced points . . 562
14.22 Finding the roots of an equation by inverse interpolation 564
14.23 The interpolation method for expanding a secular determinant . 565
*14.24 Interpolation of functions of two variables 567
*14.25 Double differences of higher order 570
*14.26 Newton's interpolation formula for a function of two variables 571
References for Chapter 14 573

CHAPTER 15
APPROXIMATE DIFFERENTIATION 574

15.1 Statement of the problem 574
15.2 Formulas of approximate differentiation based on Newton's first interpolation formula 575
15.3 Formulas of approximate differentiation based on Stirling's formula 580
15.4 Formulas of numerical differentiation for equally spaced points . 583
15.5 Graphical differentiation 586
*15.6 On the approximate calculation of partial derivatives 588
References for Chapter 15 589

CHAPTER 16
APPROXIMATE INTEGRATION OF FUNCTIONS 590

16.1 General remarks 590
16.2 Newton-Cotes quadrature formulas 593
16.3 The trapezoidal formula and its remainder term 595
16.4 Simpson's formula and its remainder term 596
16.5 Newton-Cotes formulas of higher orders 599
16.6 General trapezoidal formula (trapezoidal rule) 601
16.7 Simpson's general formula (parabolic rule) 603
16.8 On Chebyshev's quadrature formula 607
16.9 Gaussian quadrature formula 611
16.10 Some remarks on the accuracy of quadrature formulas 618
*16.11 Richardson extrapolation 622
*16.12 Bernoulli numbers 625
*16.13 Euler-Maclaurin formula 628
16.14 Approximation of improper integrals 633
16.15 The method of Kantorovich for isolating singularities 635
16.16 Graphical integration 639

*16.17 On cubature formulas 641
*16.18 A cubature formula of Simpson type 644
References for Chapter 16 648

CHAPTER 17

THE MONTE CARLO METHOD 649

17.1 The idea of the Monte Carlo method 649
17.2 Random numbers 650
17.3 Ways of generating random numbers 653
17.4 Monte Carlo evaluation of multiple integrals 656
*17.5 Solving systems of linear algebraic equations by the Monte Carlo method 666
References for Chapter 17 674
COMPLETE LIST OF REFERENCES 675
INDEX 679

INTRODUCTION

GENERAL RULES OF COMPUTATIONAL WORK

When performing computations on a large scale it is important to adhere to some simple rules that have evolved over the years and are designed to save the time and labour of the computor and make for more efficient use of computational machines and auxiliary devices.

The first step for the computor is to work out a *computational scheme* providing a detailed list of the order of operations and making it possible to achieve the desired result in the fastest and simplest manner. This is particularly necessary in computational operations of a repetitive type where a thoroughly devised scheme makes for speedy, reliable and automatic computations and fully compensates the time spent in elaborating the computational scheme. Also, a sufficiently detailed computational scheme can be competently handled by relatively inexperienced computors.

To illustrate the compilation of a computational scheme, suppose it is required to compute the values of the analytically specified function

$$y = f(x)$$

for certain values of the argument: $x = x_1, x_2, \ldots, x_n$. If the number of these values is great, it is not advisable to compute them separately, first $f(x_1)$, then $f(x_2)$, and so on, each time performing the whole sequence of operations indicated by the symbol f. It is much better to separate f into *elementary operations*

$$f(x) = f_m(\ldots(f_2(f_1(x)))\ldots)$$

and carry out the computations as repeated operations:

$$\begin{array}{ll} u_i = f_1(x_i) & (i = 1, 2, \ldots, n), \\ v_i = f_2(u_i) & (i = 1, 2, \ldots, n), \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ y = f_m(w_i) & (i = 1, 2, \ldots, n) \end{array}$$

performing one and the same operation f_j ($j=1, 2, \ldots, m$) for **all** values of the argument under consideration. Then one can make wide use of tables of functions and specialized computing devices. The results of computations should be recorded on specially designed *computing forms (computation sheets)* having appropriate divisions and headings (as applied to the chosen computational scheme). These sheets are filled in with intermediate results as they are obtained, and with the final results.

Computation sheets are usually designed so that the results of each series of repeated operations are recorded in a single column or row, and the general arrangement of intermediate results is convenient for subsequent computations.

For example, in order to compile a table of the values of the function

$$y=\frac{e^x+\cos x}{1+x^2}+\sqrt{1+\sin^2 x} \tag{1}$$

a computation sheet of the type shown in Table 1 might be recommended.

TABLE 1

COMPUTATION SHEET FOR FUNCTION (1)

x	x^2 $(1)^2$	e^x	$\sin x$	ccs x	$e^x+\cos x$ $(3)+(5)$	$1+x^2$ $(1)+(2)$	$\frac{e^x+\cos x}{1+x^2}$ $(6):(7)$	$\sin^2 x$ $(4)^2$	$1+\sin^2 x$ $(1)+(9)$	$\sqrt{1+\sin^2 x}$ $\sqrt{(10)}$	y $(8)+(11)$
1	2	3	4	5	6	7	8	9	10	11	12

The computations are performed by column. The character of the repeated operations carried out is clear from the computation sheet. First, all values of the argument x are recorded in column (1), then all the numbers of column (1) are squared and recorded in column (2). Then, using tables, the values of e^x, $\sin x$, $\cos x$ are determined in succession for each number of column (1) and are entered in columns (3), (4) and (5).

The subsequent columns give the results of intermediate operations. Say, column (6) contains the sums of $e^x+\cos x$ [schematically, $(3)+(5)$], etc. The values of the desired function y are entered in the last column, (12). With a properly constructed computation sheet, the computor does not use the given formula in his calculations since his attention is centered entirely on the sequence of filling in the indicated columns.

Note that the computational scheme and the form of the computation sheet are largely dependent on the technical facilities and auxiliary tables at hand. For example, in certain cases the separate intermediate results are stored in the memory of a machine and are not entered in the sheet. At other times, standard sets of operations may be conveniently regarded as a separate operation. For instance, in a slide-rule computation, the numerical value of an expression of the form

$$\frac{ab}{c}$$

may be computed at once without recording any intermediate result and so there is no need to split it up into the elementary operations of multiplication and division. Similarly, when using an electric desk calculator the process of finding the sum of paired products

$$\sum_{k=1}^{n} a_k b_k$$

is a single operation. In many cases it is convenient to transform the given expressions to a special artificial form (say, replace division by multiplication by the reciprocal, or bring an expression to a form convenient for taking logarithms, and so forth).

Secondly, an important step is that of *checking the computations*. A computation cannot be considered complete without a check. This stage is divided into *intermediate checks* and the *final check*. In intermediate checking, we perform certain supplementary operations that serve to convince us that the work up to a certain point has, to a certain degree of assurance, been performed correctly, otherwise the computations of that stage are repeated. The final check has to do only with the final result. For example, if the root of an equation has been computed, the value found may be substituted into the equation and we can see whether the solution is correct or not. Common sense tells us that if a computation is extensive, it is not wise to confine oneself to a final check, thus risking an enormous amount of computational effort. In such cases, it is advisable to check in stages. In responsible cases, the computations are checked by having the entire job performed by two separate computors, or the problem is done by one computor in two different ways.

A third important point to bring up here is the problem of *estimating accuracy*. In most cases, computations are carried out with approximate numbers and in approximate fashion. Therefore, even if the method of solution is exact, there will be *errors of operation* and *rounding errors* at every step in the computations. If the method is approximate, there will be added an *error of method*. Given unfavourable circumstances, the overall error may be so large that the result obtained will be purely illusory. Appropriate chapters of the book give methods for estimating errors in basic computations.

In the computation sheets, it is useful to provide columns for tabular differences (see Sec. 14.2) which may be used as a check. It is done this way. If the correctness of the difference table is violated in some section, the corresponding table entries should be recalculated (or the cause should be sought out).

One final word on the importance of making neat and legible entries in the computation sheets. Experience shows that illegible writing leads to blunders that can nullify a well organized computation. Particularly dangerous are mistakes made when writing down numbers with many zeros. Such numbers should be written in powers-of-ten notation (scientific notation, as it is sometimes called). For instance

$$0.00000345 = 3.45 \cdot 10^{-6}$$

and so on.

The rest of this book is devoted mainly to **methods of computation**. The numerical examples are in many cases simplified and intermediate operations are frequently omitted.

Chapter 1

APPROXIMATE NUMBERS

1.1 ABSOLUTE AND RELATIVE ERRORS

An *approximate number* a is a number that differs but slightly from an exact number A and is used in place of the latter in calculations. If it is known that $a < A$, then a is called a *minor* (too small) approximation of A; if $a > A$, then a is a *major* (too large) approximation of A. For example, for $\sqrt{2}$ the number 1.41 is a minor approximation while 1.42 is a major approximation, since $1.41 < \sqrt{2} < 1.42$. If a is an approximate value of the number A, we write $a \approx A$.

By the *error* Δa of an approximate number a we ordinarily mean the difference between the exact number A and the given approximate number, that is,

$$\Delta a = A - a$$

(sometimes the difference $a - A$ is called the error). If $A > a$, then the error is positive: $\Delta a > 0$; however, if $A < a$, then the error is negative: $\Delta a < 0$. To obtain the exact number A, add the error Δa to the approximate number a:

$$A = a + \Delta a$$

Thus, an exact number may be regarded as an approximate number with error zero.

In many cases the sign of the error is not known. It is then advisable to use the *absolute error of the approximate number:*

$$\Delta = |\Delta a|$$

Definition 1. The *absolute error* Δ of an approximate number a is the absolute value of the difference between the corresponding exact number A and the number a:

$$\Delta = |A - a| \tag{1}$$

Here two cases are to be distinguished:

(1) the number A is known, and then the absolute error Δ is readily determined from formula (1);

(2) the number A is not known, which is most often the case, and hence the absolute error Δ cannot be found from formula (1). It is then useful, in place of the unknown theoretical absolute error Δ, to introduce its upper estimate, the so-called *limiting absolute error*.

Definition 2. The *limiting absolute error* of an approximate number is any number not less than the absolute error of that number.

Thus, if Δ_a is the limiting absolute error of an approximate number a which takes the place of the exact number A, then

$$\Delta = |A - a| \leqslant \Delta_a \tag{2}$$

From this it follows that the exact number A lies within the range

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a \tag{3}$$

Hence $a - \Delta_a$ is a minor approximation to A, and $a + \Delta_a$ is a major approximation to A.

For brevity, we can then write

$$A = a \pm \Delta_a$$

Example 1. Determine the limiting absolute error of the number $a = 3.14$ which is used instead of the number π.

Solution. Since we have the inequality $3.14 < \pi < 3.15$ it follows that $|a - \pi| < 0.01$ and, hence, we can take $\Delta_a = 0.01$.

Taking note of the fact that

$$3.14 < \pi < 3.142$$

we have a better estimate: $\Delta_a = 0.002$.

Note that the concept of a limiting absolute error as formulated above is very broad, namely, the *limiting absolute error of an approximate number* a *is to be understood as any one of an infinity of nonnegative numbers* Δ_a *that satisfy inequality* (2). It follows logically therefrom that any number exceeding the limiting absolute error of a given approximate number can also be called the limiting absolute error of this number. For practical purposes it is convenient to take for Δ_a the smallest number (under the given circumstances) that satisfies inequality (2).

When writing an approximate number obtained from a measurement, it is common to give its limiting absolute error. For example, if the length of a line segment $l = 214$ cm to within 0.5 cm, then we write $l = 214 \text{ cm} \pm 0.5$ cm. Here the limiting absolute error $\Delta_l = 0.5$ cm, and the exact magnitude of the length l of the segment falls within the range $213.5 \text{ cm} \leqslant l \leqslant 214.5$ cm.

The absolute error (or the limiting absolute error) does not suffice to describe the accuracy of a measurement or a computation. Suppose that in measuring the lengths of two rods we get

$l_1 = 100.8$ cm ± 0.1 cm and $l_2 = 5.2$ cm ± 0.1 cm. Despite the fact that the limiting absolute errors coincide, the first measurement is better than the second one. An essential point in the accuracy of measurements is the absolute error related to unit length. It is called the *relative error*.

Definition 3. The *relative error* δ of an approximate number a is the ratio of the absolute error Δ of the number to the modulus (absolute value) of the corresponding exact number A $(A \neq 0)$. Thus

$$\delta = \frac{\Delta}{|A|} \tag{4}$$

Whence $\Delta = |A| \delta$.

As in the case of absolute errors, we introduce the notion of a *limiting relative error*.

Definition 4. The *limiting relative error* δ_a of a given approximate number a is any number not less than the relative error of that number. By definition we have

$$\delta \leqslant \delta_a \tag{5}$$

That is, $\frac{\Delta}{|A|} \leqslant \delta_a$, whence $\Delta \leqslant |A| \delta_a$.

Thus, for the limiting absolute error of a number a we can take

$$\Delta_a = |A| \delta_a \tag{6}$$

Since, in practical situations, $A \approx a$, in place of formula (6) one frequently uses

$$\Delta_a = |a| \delta_a \tag{6'}$$

From this formula, knowing the limiting relative error δ_a, we obtain the limits for the exact number. The fact that the exact number lies between $a(1-\delta_a)$ and $a(1+\delta_a)$ is symbolized as

$$A = a(1 \pm \delta_a)$$

Let a be an approximate number taking the place of an exact number A, and let Δ_a be the limiting absolute error of a. For definiteness put $A > 0$, $a > 0$ and $\Delta_a < a$. Then

$$\delta = \frac{\Delta}{A} \leqslant \frac{\Delta_a}{a - \Delta_a}$$

We can thus take the number

$$\delta_a = \frac{\Delta_a}{a - \Delta_a}$$

for the limiting relative error of the number a.

Similarly we get $\Delta = A\delta \leqslant (a+\Delta)\,\delta_a$, whence

$$\Delta_a = \frac{a\delta_a}{1-\delta_a}$$

If, as commonly occurs, $\Delta_a \ll a$ and $\delta_a \ll 1$ (the symbol $\ll$ means "very much less than"), then we can take it approximately that

$$\delta_a \approx \frac{\Delta_a}{a}$$

and

$$\Delta_a \approx a\delta_a$$

Example 2. The weight of 1 dm^3 of water at 0°C is given as $p = 999.847$ gf $\pm$ 0.001 gf (gf = gram (force)). Determine the limiting relative error of the result of weighing the water.

Solution. We obviously have $\Delta_p = 0.001$ gf and $p \leqslant 999.846$ gf. Hence

$$\delta_p = \frac{0.001}{999.846} \approx 10^{-4}\%$$

Example 3. A result of $R = 29.25$ was obtained in determining the gas constant for air. Knowing that the relative error of this value is $1^0/_{00}$ (parts per thousand), find the limits within which R lies.

Solution. We have $\delta_R = 0.001$, and so $\Delta_R = R\delta_R \approx 0.03$. Hence $29.22 \leqslant R \leqslant 29.28$.

1.2 BASIC SOURCES OF ERRORS

The errors one encounters in mathematical problems may, in the main, be broken down into five groups.

1. Errors involved in the statement of the problem. Mathematical statements rarely give an exact picture of actual phenomena. For the most part they are only idealized models. In studying the phenomena of nature we are forced, as a rule, to accept certain conditions that simplify the problem at hand. This is a source of errors (*errors of the problem*).

It sometimes happens that it is either difficult or even impossible to solve a given problem when formulated precisely. If that is the case, it is replaced by an approximate problem yielding almost the same results. This is the source of an error termed the *error of method.*

2. Errors stemming from the presence of infinite processes in mathematical analysis. The functions involved in mathematical formulas are frequently specified in the form of infinite sequences

or series $\left(\text{for example, } \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots\right)$. What is more, many mathematical equations can be solved only by describing infinite processes whose limits are the desired solutions. Since, generally speaking, an infinite process cannot be completed in a finite number of steps, we are forced to stop at some term of the sequence and consider it to be an approximation to the required solution. Naturally, such a termination of the process gives rise to an error. This error is called the *residual error*.

3. Errors due to numerical parameters (in formulas) whose values can only be determined approximately. Such, for instance, are all physical constants. Let us agree to call this error the *initial error*.

4. Errors associated with the system of numeration. When depicting even rational numbers in the decimal system or some other positional system, there may be an infinity of digits to the right of the decimal point (generally, radix point). For instance, we may have a nonterminating repeating decimal. It is obvious that we can only use a finite number of digits in our computations. This is the source of the so-called *rounding errors*. For example, assuming $\frac{1}{3} = 0.333$, we get an error of $\Delta \approx 3 \cdot 10^{-4}$. One also has to round off finite multidigit numbers.

5. Errors due to operations involving approximate numbers (*errors of operation*). When performing computations with approximate numbers, we naturally carry (to some extent) the errors of the original data into the final result. In this respect, errors of operation are **inherent**.

Quite naturally, in a specific problem some errors are absent and others exert a negligible effect. But, generally, a complete analysis must include all types of errors. In what follows we will confine ourselves largely to computing errors of operation and errors of method.

1.3 SCIENTIFIC NOTATION. SIGNIFICANT DIGITS. THE NUMBER OF CORRECT DIGITS

Any positive number a, it will be recalled, can be represented as a terminating or nonterminating decimal:

$$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \alpha_{m-2} 10^{m-2} + \dots + \alpha_{m-n+1} 10^{m-n+1} + \dots \tag{1}$$

where α_i are the digits of the number a ($\alpha_i = 0, 1, 2, \dots, 9$), the leading digit $\alpha_m \neq 0$ and m is an integer (the highest power of ten in the number a). For example,

$$3141.59\ldots = 3 \cdot 10^3 + 1 \cdot 10^2 + 4 \cdot 10^1 + 1 \cdot 10^0 + \\ + 5 \cdot 10^{-1} + 9 \cdot 10^{-2} + \dots$$

Each unit occupies a specific position in the number a written as the decimal fraction (1) and has a definite value. The unit standing in the first position is equal to 10^m, that in the second position, 10^{m-1}, in the nth position, 10^{m-n+1}, etc.

Actual cases usually involve approximate numbers in the form of **terminating** decimals:

$$b=\beta_m 10^m+\beta_{m-1}10^{m-1}+\ldots+\beta_{m-n+1}10^{m-n+1} \qquad (\beta_m\neq 0) \qquad (2)$$

All retained decimal digits β_i $(i=m,\ m-1,\ \ldots,\ m-n+1)$ are called *significant digits* of the approximate number b; note that some of them may be equal to zero (with the exception of β_m). In the decimal positional system of representing the number b, one often has to introduce zeros at the beginning or the end of the number. To illustrate,

$$b=7\cdot 10^{-3}+0\cdot 10^{-4}+1\cdot 10^{-5}+0\cdot 10^{-6}=\underline{0.00}7010$$

or

$$b=2\cdot 10^{9}+0\cdot 10^{8}+0\cdot 10^{7}+3\cdot 10^{6}+0\cdot 10^{5}=2{,}003{,}0\underline{00{,}000}$$

The underlined zeros are not significant digits.

Definition 1. A *significant digit* of an approximate number is any nonzero digit, in its decimal representation, or any zero lying between significant digits or used as a placeholder, to indicate a retained place. All other zeros of the approximate number that serve only to fix the position of the decimal point are not to be considered significant digits.

For example, in the number 0.002080 the first three zeros are not significant digits since they serve only to fix the position of the decimal point and indicate the place values of the other digits. The other two zeros are significant digits since the first lies between the digits 2 and 8 and the second (as indicated by the notation) shows that we retain the decimal place 10^{-6} in the approximate number. If the last digit of 0.002080 is not significant, then the number must be written as 0.00208. From this point of view, the numbers 0.002080 and 0.00208 are not the same, because the former has four significant digits and the latter only three.

When writing large numbers, the zeros on the right can serve both to indicate the significant digits and to fix the place values of the other digits. This can lead to misunderstanding when the numbers are written in the ordinary way. Consider, for instance, the number 689,000. It is not clear how many significant digits there are, although we can say we have at least three. This ambiguity can be avoided by using powers-of-ten notation (scientific notation) and writing the number as $6.89\cdot 10^5$ if it has three significant digits, or as $6.8900\cdot 10^5$ if the number has five significant digits, etc. Generally speaking, this notation is convenient for numbers containing a large number of nonsignificant zeros, such as $0.000000120 = 1.20 \cdot 10^{-7}$, and the like.

Let us introduce the notion of *correct digits* of an approximate number.

Definition 2. We say that the first n significant digits of an approximate number are *correct* if the absolute error of the number does not exceed one half unit in the nth place, counting from left to right.

Thus, if for an approximate number a (1), which takes the place of an exact number A, it is known that

$$\Delta = |A - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1}$$

then by definition the first n digits $\alpha_m, \alpha_{m-1}, \ldots, \alpha_{m-n+1}$ of this number are correct.

For example, with respect to the exact number $A = 35.97$, the number $a = 36.00$ is an approximation correct to three digits, since $|A - a| = 0.03 < \frac{1}{2} \cdot 0.1$.

Note that in mathematical tables all indicated significant digits are correct. For instance, in a five-place table of logarithms the absolute error of a mantissa definitely does not exceed $\frac{1}{2} \cdot 10^{-5}$, etc.

The term "n correct digits" is not to be taken literally; that is, it is not necessarily true that in a given approximate number a having n true digits, the first n significant digits of it coincide with the corresponding digits of the exact number A. For example, the approximate number $a = 9.995$, which stands for the exact number $A = 10$, is correct to three digits, yet all the digits of these numbers differ. However, in many cases we do find that the correct digits of the approximate number are the same as the corresponding digits of the exact number.

Note. In some cases it is convenient to say that the number a is an approximation to an exact number A to n *correct digits* in the broad sense, meaning by this that the absolute error $\Delta = |A - a|$ does not exceed one unit in the nth significant digit of the approximate number.

For example, with respect to the exact number $A = 412.3567$, the number $a = 412.356$ is an approximation correct to six digits in the broad sense, since $\Delta = 0.0007 < 1 \cdot 10^{-3}$.

In the sequel we will regard the correct digits of an approximate number in the sense of Definition 2 (that is to say, *in the narrow sense*), unless otherwise stated.

1.4 ROUNDING OF NUMBERS

Consider an approximate or exact number a written in the decimal number system. It is often required to round off this number, which is to say, to replace it by a number a_1 having a smaller number of significant digits. The number a_1 is chosen so as to keep the *rounding error* $|a_1-a|$ to a minimum.

Rounding-off rule. In order to round off a number to n significant digits, drop all digits to the right of the nth significant digit, or replace them by zeros if the zeros are needed as placeholders. In this operation, note the following:

(1) if the first of the discarded digits is less than 5, leave the remaining digits unchanged;

(2) if the first discarded digit exceeds 5, add 1 to the last retained digit;

(3) if the first discarded digit is exactly 5 and there are nonzero digits among those discarded, add unity to the last retained digit;

(3a) however, if the first discarded digit is exactly 5 and all other discarded digits are zeros, the last retained digit is left unchanged if even and is increased by unity if odd (the **even-digit rule**).

In other words, if in rounding off a number we discard less than half a unit of the last retained digit, all the retained digits are left unaltered; but if the discarded portion of the number is more than one half unit of the last retained place, the digit of that place is increased by unity. In the exceptional case when the discarded part is **exactly** equal to one-half unit of the last retained place, the even-digit rule is invoked in order to compensate for the signs of errors due to rounding.

It is obvious that when applying the rounding-off rule, the rounding error does not exceed one half unit in the place of the last retained significant digit.

Example 1. Rounding the number

$$\pi = 3.1415926535\ldots$$

to five, four and three significant digits, we get the approximate numbers 3.1416, 3.142, 3.14 with absolute errors less than $\frac{1}{2}\cdot 10^{-4}$, $\frac{1}{2}\cdot 10^{-3}$ and $\frac{1}{2}\cdot 10^{-2}$.

Example 2. Rounding the number 1.2500 to two significant digits, we obtain the approximate number 1.2 with an absolute error equal to $\frac{1}{2}\cdot 10^{-1}=0.05$.

The accuracy of an approximate number does not depend on the number of significant digits, but on the number of **correct significant digits** [1], [2]. When an approximate number contains extra incorrect significant digits, one resorts to rounding. The following **practical rule** may be used as a guide: *when performing approximate computations, the number of significant digits in the intermediate results must not exceed the number of correct digits by more than one or two units.* The final result must not contain more than one extra significant digit over the number of correct digits. If the absolute error of the result does not exceed two units of the last retained place, the extra digit is *in doubt.*

The foregoing rule makes for a large saving in time (by dispensing with extra digits) without impairing the accuracy of the computation. Retention of additional digits has the meaning that ordinarily an error estimate of the results is made relative to the worst version, and the actual error may be appreciably less than the maximum theoretical error. Thus, in many cases, significant digits that are considered incorrect are actually correct.

One also rounds off exact numbers that contain either too many or an infinity of significant digits, depending on the general required accuracy of the computations.

Note that if an exact number A is rounded off to n significant digits by the rounding-off rule, the resulting approximate number a will have n correct digits (in the narrow sense).

Now if an approximate number a having n correct digits is rounded off to n significant digits, the resulting new approximate number a_1 will, generally speaking, have n correct digits in the broad sense. Indeed, by virtue of the inequality

$$|A-a_1| \leqslant |A-a|+|a-a_1|$$

the limiting absolute error of the number a_1 is made up of the absolute error of a and the rounding error.

1.5 RELATIONSHIP BETWEEN THE RELATIVE ERROR OF AN APPROXIMATE NUMBER AND THE NUMBER OF CORRECT DIGITS

We will now prove a theorem that relates the magnitude of the relative error of an approximate number to the number of correct digits in that number [3], [4].

Theorem. *If a positive approximate number* a *has* n *correct digits in the narrow sense, the relative error* δ *of this number does not exceed* $\left(\frac{1}{10}\right)^{n-1}$ *divided by the first significant digit of the given* *number, or*

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}$$

where α_m *is the first significant digit of number* a.

Proof. Let the number

$$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \ldots + \alpha_{m-n+1} 10^{m-n+1} + \ldots \qquad (\alpha_m \geqslant 1)$$

be an approximate value of the exact number A and let it be correct to n digits. By definition we then have

$$\Delta = |A - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1}$$

whence

$$A \geqslant a - \frac{1}{2} \cdot 10^{m-n+1}$$

This inequality is further strengthened if the number a is replaced by a definitely smaller number $\alpha_m 10^m$:

$$A \geqslant \alpha_m 10^m - \frac{1}{2} \cdot 10^{m-n+1} = \frac{1}{2} \cdot 10^m \left(2\alpha_m - \frac{1}{10^{n-1}}\right) \qquad (1)$$

The right side of inequality (1) is a minimum for $n = 1$. Therefore

$$A \geqslant \frac{1}{2} \cdot 10^m (2\alpha_m - 1) \qquad (2)$$

or, since

$$2\alpha_m - 1 = \alpha_m + (\alpha_m - 1) \geqslant \alpha_m$$

it follows that

$$A \geqslant \frac{1}{2} \alpha_m 10^m$$

Hence

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2} 10^{m-n+1}}{\frac{1}{2} \alpha_m 10^m} = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}$$

Thus

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1} \qquad (3)$$

and the theorem is proved.

Note 1. Inequality (2) may be used to obtain a more exact estimate of the relative error δ.

Corollary 1. For the limiting relative error of the number a we can take

$$\delta_a = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1} \tag{4}$$

where α_m is the first significant digit of the number a.

Corollary 2. If the number a has more than two correct digits, that is, $n \geqslant 2$, then for all practical purposes the following formula holds:

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1} \tag{5}$$

Indeed, for $n \geqslant 2$ we can neglect $\frac{1}{10^{n-1}}$ in inequality (1). Then

$$A \geqslant \frac{1}{2} \cdot 10^m \cdot 2\alpha_m = \alpha_m\, 10^m$$

whence

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2} \cdot 10^{m-n+1}}{\alpha_m\, 10^m} = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}$$

Consequently

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}$$

Note 2. If the approximate number a is correct to n digits in the broad sense, (4) and (5) should be increased by a factor of 2.

Example 1. What is the limiting relative error if we take $a = 3.14$ in place of the number π?

Solution. In our case $\alpha_m = 3$ and $n = 3$, and so

$$\delta_a = \frac{1}{2 \cdot 3}\left(\frac{1}{10}\right)^{3-1} = \frac{1}{600} = \frac{1}{6}\ \%$$

Example 2. How many digits are to be taken in computing $\sqrt{20}$ so that the error does not exceed 0.1%?

Solution. Since the first digit is 4, we have $\alpha_m = 4$ and $\delta = 0.001$. We have $\frac{1}{4 \cdot 10^{n-1}} \leqslant 0.001$, whence $10^{n-1} \geqslant 250$ and $n \geqslant 4$.

This theorem enables us to determine the relative error δ of an approximate number a from the number of correct digits:

$$a = \alpha_m\, 10^m + \alpha_{m-1}\, 10^{m-1} + \ldots \tag{6}$$

To solve the inverse problem—to determine the number of n correct digits of number (6) if the relative error δ is known—we ordinarily use the approximate formula

$$\delta = \frac{\Delta}{a} \quad (a > 0)$$

where Δ is the absolute error of a. Whence

$$\Delta = a\delta \tag{7}$$

Taking into account the leading power of ten in the number Δ, it is easy to establish the number of correct digits in the given approximate number a. In particular, if

$$\delta \leqslant \frac{1}{10^n}$$

then from formulas (6) and (7) we have

$$\Delta \leqslant (\alpha_m + 1)\, 10^m \cdot 10^{-n} \leqslant 10^{m-n+1}$$

In other words, a is definitely correct to n decimal places in the broad sense. Similarly, if

$$\delta \leqslant \frac{1}{2 \cdot 10^n}$$

then the number a is correct to n places in the narrow sense.

Example 3. An approximate number $a = 24{,}253$ has a relative error of 1%. How many correct digits has it?

Solution. We have

$$\Delta = 24{,}253 \cdot 0.01 \approx 243 = 2.43 \cdot 10^2$$

Thus, the number a has only two correct digits ($n = 2$); the hundreds digit is in doubt. According to the rule given above, the number a is preferably written as $a = 2.43 \cdot 10^4$.

Note. The foregoing method for determining the number of correct digits is approximate. In an exact count of the correct digits of number a, one should proceed from the inequalities

$$\delta \geqslant \frac{\Delta}{a + \Delta}$$

and

$$\Delta \leqslant \frac{a\delta}{1-\delta} \quad (0 \leqslant \delta < 1)$$

1.6 TABLES FOR DETERMINING THE LIMITING RELATIVE ERROR FROM THE NUMBER OF CORRECT DIGITS AND VICE VERSA

If an approximate number is written with indicated correct digits, then it is easy to compute its limiting relative error. This is a frequent requirement in practical computations and so it is desirable to simplify the operation. Table 2 [5] indicates the relative error as a percentage of the approximate number depending on the number of correct digits (in the broad sense) and on the first two significant digits of the number, counting from left to right.

TABLE 2

RELATIVE ERROR (IN %) OF NUMBERS CORRECT TO n DIGITS

First two significant digits	n		
	2	3	4.
10-11	10	1	0.1
12-13	8.3	0.83	0.083
14, ..., 16	7.1	0.71	0.071
17, ..., 19	5.9	0.59	0.059
20, ..., 22	5	0.5	0.05
23, ..., 25	4.3	0.43	0.043
26, ..., 29	3.8	0.38	0.038
30, ..., 34	3.3	0.33	0.033
35, ..., 39	2.9	0.29	0.029
40, ..., 44	2.5	0.25	0.025
45, ..., 49	2.2	0.22	0.022
50, ..., 59	2	0.2	0.02
60, ..., 69	1.7	0.17	0.017
70, ..., 79	1.4	0.14	0.014
80, ..., 89	1.2	0.12	0.012
90, ..., 99	1.1	0.11	0.011

By way of an example, suppose we have an approximate number 0.00354 correct to three decimals. Since $n=3$ and the number 35 lies in the interval 35, ..., 39, from Table 2 we find $\delta=0.29\%$.

If only the first digit of the number is known, say 4, then of course we take the greater of the numbers 2.5 and 2.2, which correspond to possible versions 40, ..., 44 and 45, ..., 49 (for $n=2$). If the first digit is unknown, then we take the numbers from the first row (10%, 1%, and 0.1%) as the largest ones. From this table we see that three correct digits ensure a relative accuracy (with an error not exceeding 1%) sufficient for most computations. Note that if the approximate number is correct to two, three or four digits in the narrow sense, then all the numbers of the table must be halved.

Table 3 [5] gives upper bounds for relative errors (in percent) that ensure a given approximate value a certain number of correct digits in the broad sense depending on its first two digits.

TABLE 3

NUMBER OF CORRECT DIGITS OF AN APPROXIMATE NUMBER DEPENDING ON THE LIMITING RELATIVE ERROR (IN %)

First two significant digits	n		
	2	3	4
10-11	4.2	0.42	0.042
12-13	3.6	0.36	0.036
14, ..., 16	2.9	0.29	0.029
17, ..., 19	2.5	0.25	0.025
20, ..., 22	2.2	0.22	0.022
23, ..., 25	1.9	0.19	0.019
26, ..., 29	1.7	0.17	0.017
30, ..., 34	1.4	0.14	0.014
35, ..., 39	1.2	0.12	0.012
40, ..., 44	1.1	0.11	0.011
45, ..., 49	1	0.1	0.01
50, ..., 54	0.9	0.09	0.009
55, ..., 59	0.8	0.08	0.008
60, ..., 69	0.7	0.07	0.007
70, ..., 79	0.6	0.06	0.006
80, ..., 99	0.5	0.05	0.005

An example will serve to illustrate how to use Table 3. Suppose we have an approximate number $a = 5.297$ with relative error $\delta = 0.5\%$. The first two significant digits here are 5 and 2; the number made up of these digits lies between 50 and 54. Depending on the number of correct digits, the latter are associated with relative errors of 0.9%, 0.09% and 0.009%, etc. Since $\delta = 0.5\% < 0.9\%$ and the relative error of a number does not depend on what decimal places are expressed by the digits of the number, the number $a = 5.297$ is correct to two decimals in the broad sense.

Example 1. Taking $\pi = 3.142$, $\sqrt{7} = 2.65$, $e = 2.718$, $\log_{10} 5 = 0.699$, $\sin 1^\circ = 0.0174$, we find from Table 2 that the corresponding relative errors are: $\delta = 0.033\%$, $\delta = 0.19\%$, $\delta = 0.019\%$, $\delta = 0.17\%$, $\delta = 0.59\%$.

Example 2. From the deflection of a steel rod, Young's modulus has been computed as $E = 2212\ldots$ T/cm^2 (T = metric tons) to within 2%. How many digits are correct in the value found? From Table 3 we find $n = 2$, and so $E = 22 \cdot 10^2$ T/cm^2.

Example 3. The gas constant R is computed for the explosive mixture used in a gas motor. $R = 31.5\ldots$ with a relative error of $\delta = 1\%$. Find the number of correct digits. From Table 3 we have $n = 2$, and so $R = 32$.

1.7 THE ERROR OF A SUM

Theorem 1. *The absolute error of an algebraic sum of several approximate numbers does not exceed the sum of the absolute errors of the numbers.*

Proof. Let $x_1, x_2, \ldots, x_n$ be the given approximate numbers. Consider their algebraic sum

$$u = \pm x_1 \pm x_2 \pm \ldots \pm x_n$$

Obviously

$$\Delta u = \pm \Delta x_1 \pm \Delta x_2 \pm \ldots \pm \Delta x_n$$

and, hence,

$$|\Delta u| \leqslant |\Delta x_1| + |\Delta x_2| + \ldots + |\Delta x_n| \tag{1}$$

Corollary. For the limiting absolute error of an algebraic sum we can take the sum of the limiting absolute errors of the terms:

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n} \tag{2}$$

From (2) it follows that the limiting absolute error of a sum cannot be less than the limiting absolute error of the least accurate term (in the sense of the absolute error), which is to say the term having the maximum absolute error. Consequently, no matter how high the degree of accuracy of the other terms, we cannot, through them, increase the accuracy of the sum. For this reason, it is meaningless to retain extra digits in the more exact terms. From the foregoing we obtain the following practical rule for the addition of approximate numbers.

Rule. To add numbers of different absolute accuracy,

(1) find the numbers with the least number of decimal places and leave them unchanged;

(2) round off the remaining numbers, retaining one or two more decimal places than those with the smallest number of decimals;

(3) add the numbers, taking into account all retained decimals;

(4) round off the result, reducing it by one decimal.

When rounding to the mth place the terms of a sum by the rounding-off rule,

$$u = x_1 + x_2 + \ldots + x_n$$

the rounding error of the sum does not exceed

$$\Delta_{\text{round}} \leqslant n \cdot \frac{1}{2} \cdot 10^m \tag{3}$$

in the most unfavourable case.

A more exact calculation of the rounding error of a sum may be obtained if we take into account the signs of the rounding errors of the terms.

Example. Find the sum of the approximate numbers 0.348, 0.1834, 345.4, 235.2, 11.75, 9.27, 0.0849, 0.0214, 0.000354, each correct to the indicated significant digits (in the broad sense).

Solution. We find the least accurate numbers 345.4 and 235.2 whose absolute error may attain 0.1. Rounding the remaining numbers to 0.01, we get

$$\begin{array}{r}
345.4 \\
235.2 \\
11.75 \\
9.27 \\
0.35 \\
0.18 \\
0.08 \\
0.02 \\
0.00 \\
\hline
602.25
\end{array}$$

Rounding the result to 0.1 by the even-digit rule, we get the approximate value of the sum: 602.2.

The total error Δ of the result is made up of three terms:

(1) the sum of the limiting errors of the original data:

$$\Delta_1 = 10^{-3} + 10^{-4} + 10^{-1} + 10^{-1} + 10^{-2} + 10^{-2} + 10^{-4} + 10^{-4} + 10^{-6} = \\ = 0.221301 < 0.222$$

(2) the absolute value of the sum of the rounding errors of the terms (with regard for signs)

$$\Delta_2 = |-0.002 + 0.0034 + 0.0049 + 0.0014 + 0.000354| = \\ = 0.008054 < 0.009$$

(3) the final rounding error of the result:

$$\Delta_3 = 0.050$$

Hence

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3 \leqslant 0.222 + 0.009 + 0.050 = 0.281 < 0.3$$

and thus the desired sum is 602.2 ± 0.3.

Theorem 2. *If the terms have one and the same sign, the limiting relative error of their sum does not exceed the maximum limiting relative error of any of the terms.*

Proof. Let $u = x_1 + x_2 + \ldots + x_n$, where, for definiteness, $x_i > 0$ $(i = 1, 2, \ldots, n)$.

Denote by A_i $(A_i > 0;\ i = 1, 2, \ldots, n)$ the exact magnitudes of the terms x_i, and by $A = A_1 + A_2 + \ldots + A_n$ the exact value of the sum u. Then, for the limiting relative error of the sum we can take

$$\delta_u = \frac{\Delta_u}{A} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{A_1 + A_2 + \ldots + A_n} \tag{4}$$

Since

$$\delta_{x_i} = \frac{\Delta_{x_i}}{A_i} \quad (i = 1, 2, \ldots, n)$$

it follows that

$$\Delta_{x_i} = A_i \delta_{x_i} \tag{4'}$$

Substituting this expression into (4), we get

$$\delta_u = \frac{A_1 \delta_{x_1} + A_2 \delta_{x_2} + \ldots + A_n \delta_{x_n}}{A_1 + A_2 + \ldots + A_n}$$

Let $\bar{\delta}$ be the greatest of the relative errors δ_{x_i}, or $\delta_{x_i} \leqslant \bar{\delta}$. Then

$$\delta_u \leqslant \frac{\bar{\delta}(A_1 + A_2 + \ldots + A_n)}{A_1 + A_2 + \ldots + A_n} = \bar{\delta}$$

Consequently, $\delta_u \leqslant \bar{\delta}$, or

$$\delta_u \leqslant \max(\delta_{x_1}, \delta_{x_2}, \ldots, \delta_{x_n})$$

1.8 THE ERROR OF A DIFFERENCE

We consider the difference of two approximate numbers: $u = x_1 - x_2$. From formula (2) of Sec. 1.7, the limiting absolute error Δ_u of the difference is

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2}$$

That is, *the limiting absolute error of a difference is equal to the sum of the limiting absolute errors of the diminuend and the subtrahend.*

Whence the limiting relative error of the difference is

$$\delta_u = \frac{\Delta_{x_1} + \Delta_{x_2}}{A} \tag{1}$$

where A is the exact value of the absolute magnitude of the difference between the numbers x_1 and x_2.

Note on the loss of accuracy when subtracting nearly equal numbers. If the approximate numbers x_1 and x_2 are nearly equal numbers and have small absolute errors, the number A is small. From formula (1) it follows that the limiting relative error in this case can be very great whereas the relative errors of the diminuend and subtrahend remain small. This amounts to a *loss of accuracy*.

To illustrate let us compute the difference between two numbers: $x_1 = 47.132$ and $x_2 = 47.111$, each of which is correct to five significant digits. Subtracting, we get $u = 47.132 - 47.111 = 0.021$.

The difference u has only two significant digits, of which the last is uncertain since the limiting absolute error of the difference is

$$\Delta_u = 0.0005 + 0.0005 = 0.001$$

The limiting relative errors of the subtrahend, diminuend, and difference are:

$$\delta_{x_1} = \frac{0.0005}{47.132} \approx 0.00001,$$

$$\delta_{x_2} = \frac{0.0005}{47.111} \approx 0.00001,$$

$$\delta_u = \frac{0.001}{0.021} \approx 0.05$$

The limiting relative error of the difference is roughly 5000 times greater than the limiting relative errors of the original figures.

It is therefore desirable, in approximate computations, to transform the expressions in which computation of numerical values leads to the subtraction of nearly equal numbers.

Example. Find the difference

$$u = \sqrt{2.01} - \sqrt{2} \tag{2}$$

to three correct digits.

Solution. Since

$$\sqrt{2.01} = 1.41774469\ldots$$

and

$$\sqrt{2} = 1.41421356\ldots$$

the desired result is

$$u = 0.00353 = 3.53 \cdot 10^{-3}$$

This result can be obtained by writing expression (2) as

$$u = \frac{0.01}{\sqrt{2.01} + \sqrt{2}}$$

and then finding the roots of $\sqrt{2.01}$ and $\sqrt{2}$ to three correct digits, as witness,

$$u = \frac{0.01}{1.42 + 1.41} = \frac{0.01}{2.83} = 10^{-2} \cdot 3.53 \cdot 10^{-1} = 3.53 \cdot 10^{-3}$$

From what has been said we can formulate a practical rule: in approximate computations avoid as far as possible the subtraction of two nearly equal approximate numbers; if it is necessary to subtract such numbers, take the diminuend and subtrahend with a sufficient number of additional correct digits (if that is possible). For example, if we desire the difference of two numbers x_1 and x_2 to n

significant digits and it is known that the first m significant digits will disappear by subtraction, then we must start with $m+n$ significant digits in each of the numbers (x_1 and x_2).

1.9 THE ERROR OF A PRODUCT

Theorem. *The relative error of a product of several approximate nonzero numbers does not exceed the sum of the relative errors of the numbers.*

Proof. Let $u = x_1 x_2 \ldots x_n$.

Assuming for the sake of simplicity that the approximate numbers $x_1, x_2, \ldots, x_n$ are positive, we have

$$\ln u = \ln x_1 + \ln x_2 + \ldots + \ln x_n$$

Whence, using the approximate formula $\Delta \ln x \approx d \ln x = \frac{\Delta x}{x}$, we get

$$\frac{\Delta u}{u} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} + \ldots + \frac{\Delta x_n}{x_n}$$

Taking the absolute value of the latter expression, we obtain

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right| + \ldots + \left|\frac{\Delta x_n}{x_n}\right|$$

If $A_i\,(i = 1, 2, \ldots, n)$ are the exact values of the factors x_i, and $|\Delta x_i|$, as is usually the case, are small compared with x_i, we can approximately set

$$\left|\frac{\Delta x_i}{x_i}\right| \approx \left|\frac{\Delta x_i}{A_i}\right| = \delta_i$$

and

$$\left|\frac{\Delta u}{u}\right| = \delta$$

where δ_i are the relative errors of the factors $x_i\,(i = 1, 2, \ldots, n)$ and δ is the relative error of the product.

Consequently

$$\delta \leqslant \delta_1 + \delta_2 + \ldots + \delta_n \qquad (1)$$

It is obvious that (1) also holds true if the factors x_i $(i = 1, 2, \ldots, n)$ have different signs.

Corollary. The limiting relative error of a product is equal to the sum of the limiting relative errors of the factors:

$$\delta_u = \delta_{x_1} + \delta_{x_2} + \ldots + \delta_{x_n} \qquad (2)$$

If all factors of the product u are extremely exact, with the exception of one, then from (2) it follows that the limiting relative error of the product will practically coincide with the limiting relative error of the least accurate factor. In the particular case of only the factor x_1 being approximate, we simply have

$$\delta_u = \delta_{x_1}$$

Knowing the limiting relative error δ_u of a product u, we can determine its limiting absolute error Δ_u by the formula

$$\Delta_u = |u| \delta_u$$

Example 1. Determine the product u of the approximate numbers $x_1 = 12.2$ and $x_2 = 73.56$ and the number of correct digits in it if the factors are correct to all written digits.

Solution. We have $\Delta_{x_1} = 0.05$ and $\Delta_{x_2} = 0.005$, whence

$$\delta_u = \frac{0.05}{12.2} + \frac{0.005}{73.56} = 0.0042$$

Since the product $u = 897.432$, it follows that $\Delta_u = u\delta_u = 897 \cdot 0.004 = 3.6$ (approximately).

And so u is correct to only two digits and the result should be written as

$$u = 897 \pm 4$$

Note the special case

$$u = kx$$

where k is an exact factor different from zero. We have

$$\delta_u = \delta_x$$

and

$$\Delta_u = |k| \Delta_x$$

When multiplying an approximate number by an exact factor k, *the limiting relative error remains unchanged, while the limiting absolute error is increased* |k|*-fold.*

Example 2. In aiming a rocket at a target, the limiting angular error is $\varepsilon = 1'$. What is the possible deviation Δ_u of the rocket from the target over a range of $x = 2{,}000$ km in the absence of error correction?

Solution. Here

$$\Delta_u = \frac{\pi}{180 \cdot 60} \cdot 2{,}000 \text{ km} \approx 580 \text{ m}$$

It is clear that the relative error of a product cannot be less than the relative error of the least accurate factor. For this reason, as in addition, it is meaningless to retain extra significant digits in the more precise factors.

The following rule is a useful guide: in order to find the product of several approximate numbers correct to different numbers of significant digits, it is sufficient to:

(1) round them off so that each contains one or two significant digits more than the number of correct digits in the least accurate factor;

(2) in the final result retain as many significant digits as there are correct digits in the least accurate factor (or keep one extra digit).

Example 3. Find the product of the approximate numbers $x_1 = 2.5$ and $x_2 = 72.397$ correct to the number of digits written.

Solution. Using the rule, we have, after rounding, $x_1 = 2.5$ and $x_2 = 72.4$, whence $x_1 x_2 = 2.5 \cdot 72.4 = 181 \approx 1.8 \cdot 10^2$.

1.10 THE NUMBER OF CORRECT DIGITS IN A PRODUCT

Suppose we have a product of n factors $(n \leqslant 10)$ $u = x_1 x_2 \ldots x_n$, each correct to at least m $(m > 1)$ digits. Assume also that α_1, α_2, $\ldots$, α_n are the first significant digits in a decimal representation of the factors:

$$x_i = \alpha_i 10^{p_i} + \beta_i 10^{p_i - 1} + \ldots \quad (i = 1, 2, 3, \ldots, n)$$

Then by formula (5), Sec. 1.5, we have

$$\delta_{x_i} = \frac{1}{2\alpha_i} \left(\frac{1}{10} \right)^{m-1} \qquad (i = 1, 2, \ldots, n)$$

and, hence,

$$\delta_u = \frac{1}{2} \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n} \right) \left(\frac{1}{10} \right)^{m-1} \tag{1}$$

Since $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n} \leqslant 10$, it follows that $\delta_u \leqslant \frac{1}{2} \left(\frac{1}{10} \right)^{m-2}$.

Consequently the product u is correct to $m-2$ digits in the most unfavourable case.

Rule. If all factors are correct to m decimal places and their number does not exceed 10, then the number of correct (in the broad sense) digits of the product is less than m by one or two units.

Consequently if m correct decimal places are required in a product, the factors should be taken with one or two extra digits.

If the factors are of different accuracies, then m is to mean the number of correct digits in the least accurate factor. Thus, *the number of correct digits in a product of a small number of factors (of the order of ten) may be one or two units less than the number of correct digits in the least accurate factor.*

Example 1. Determine the relative error and the number of correct digits in the product $u=93.87\cdot 9.236$.

Solution. By formula (1) we have

$$\delta_u=\frac{1}{2}\left(\frac{1}{9}+\frac{1}{9}\right)\frac{1}{10^3}=\frac{1}{9}\cdot 10^{-3}<\frac{1}{2}\cdot 10^{-3}$$

Hence the product u is correct to at least three digits (see Sec. 1.5).

Example 2. Determine the relative error and the number of correct digits in the product $u=17.63\cdot 14.285$.

Solution.

$$\delta_u=\frac{1}{2}\left(\frac{1}{1}+\frac{1}{1}\right)\frac{1}{10^3}=1\cdot 10^{-3}$$

Thus the product will contain at least three correct digits (in the broad sense).

1.11 THE ERROR OF A QUOTIENT

If $u=\frac{x}{y}$, then $\ln u=\ln x-\ln y$

and

$$\frac{\Delta u}{u}=\frac{\Delta x}{x}-\frac{\Delta y}{y}$$

whence

$$\left|\frac{\Delta u}{u}\right|\leqslant\left|\frac{\Delta x}{x}\right|+\left|\frac{\Delta y}{y}\right|$$

This formula shows that the theorem of Sec. 1.9 holds true for a quotient as well.

Theorem. *The relative error of a quotient does not exceed the sum of the relative errors of the dividend and divisor.*

Corollary. If $u=\frac{x}{y}$, then $\delta_u=\delta_x+\delta_y$.

Example. Find the number of correct digits in the quotient $u=25.7:3.6$ assuming the dividend and divisor are correct to the last digit given.

Solution. We have

$$\delta_u=\frac{0.05}{25.7}+\frac{0.05}{3.6}=0.002+0.014=0.016$$

Since $u=7.14$, then $\Delta_u=0.016\cdot 7.14=0.11$. And so the quotient u is correct to two digits in the broad sense, that is, $u=7.1$ or, more precisely,

$$u=7.14\pm 0.11$$

1.12 THE NUMBER OF CORRECT DIGITS IN A QUOTIENT

Suppose the dividend x and the divisor y are correct to at least m digits. If α and β are their first significant digits, then for the limiting relative error of the quotient u we can take the quantity

$$\delta_u = \frac{1}{2}\left(\frac{1}{\alpha}+\frac{1}{\beta}\right)\left(\frac{1}{10}\right)^{m-1}$$

From this we get the rule: (1) if $\alpha \geqslant 2$ and $\beta \geqslant 2$, then the quotient u is correct to at least $m-1$ digits; (2) if $\alpha = 1$ or $\beta = 1$, then the quotient u is definitely correct to $m-2$ digits.

1.13 THE RELATIVE ERROR OF A POWER

Suppose $u = x^m$ (m any natural number), then $\ln u = m \ln x$ and, hence,

$$\left|\frac{\Delta u}{u}\right| = m\left|\frac{\Delta x}{x}\right|$$

From this we have

$$\delta_u = m\delta_x \tag{1}$$

or *the limiting relative error of the* m*th power of a number is* m *times the limiting relative error of the number.*

1.14 THE RELATIVE ERROR OF A ROOT

Now suppose $u = \sqrt[m]{x}$; then $u^m = x$, whence

$$\delta_u = \frac{1}{m}\delta_x \tag{1}$$

That is, *the limiting relative error of a root of index* m *is* m *times less than the limiting relative error of the radicand.*

Example. Determine with what relative error and with how many correct digits we can find the side a of a square if its area $s = 12.34$ to the nearest hundredth.

Solution. We have $a = \sqrt{s} = 3.5128\ldots$. Since

$$\delta_s = \frac{0.01}{12.33} \approx 0.0008$$

it follows that $\delta_a = \frac{1}{2}\delta_s = 0.0004$. Therefore

$$\Delta_a = 3.5128 \cdot 0.0004 = 1.4 \cdot 10^{-3}$$

And so the number a will have about four correct digits (in the broad sense) and, hence, $a = 3.513$.

1.15 COMPUTATIONS IN WHICH ERRORS ARE NOT TAKEN INTO EXACT ACCOUNT

In the preceding sections we indicated ways of estimating the limiting absolute error of operations. It was assumed that the absolute errors of the components strengthen one another, but actually this is rarely the case.

In large-scale computations, when the error of each separate result is not taken into account, it is advisable to apply the following rules for counting digits [6].

1. When adding and subtracting approximate numbers, the last retained digit in the result must be the largest of the decimal orders expressed by the last correct significant digits of the original data.
2. When multiplying and dividing approximate numbers, retain in the result as many significant digits as there are in the given approximate number having the smallest number of correct significant digits.
3. When squaring or cubing an approximate number, retain as many significant digits in the result as there are correct significant digits in the base of the power.
4. When taking square or cube roots of an approximate number, take as many significant digits in the answer as there are correct digits in the radicand.
5. In all intermediate results, keep one more digit than recommended by the rules given earlier. Discard this additional digit in the final result.
6. When using logarithms, count the number of correct significant digits in the approximate number having the smallest number of correct significant digits and use tables of logarithms to one place more than that number. The last significant digit is discarded in the answer.
7. If the data can be taken with arbitrary accuracy, then in order to obtain a result correct to k digits, take the original data with the number of digits such that according to the earlier rules we obtain $k+1$ correct digits in the answer.

If some of the data have extra lower-order digits (in addition and subtraction) or more significant digits than the others (in multiplication, division, powers or roots), then they must first be rounded off so that one additional digit is retained.

1.16 GENERAL FORMULA FOR ERRORS

The prime objective of the theory of errors is: given the errors of a certain set of quantities, to determine the error of a given function of these quantities.

Suppose a differentiable function

$$u = f(x_1, x_2, \ldots, x_n)$$

is given; let $|\Delta x_i|$ $(i = 1, 2, \ldots, n)$ be the absolute errors of the arguments of the function. Then the absolute error of the function is

$$|\Delta u| = |f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n)|$$

In practical situations, $|\Delta x_i|$ are ordinarily small quantities whose products, squares and higher powers may be neglected. We can therefore put

$$|\Delta u| \approx |df(x_1, x_2, \ldots, x_n)| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i\right| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i|$$

Thus

$$|\Delta u| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i| \tag{1}$$

From this, denoting by Δ_{x_i} $(i = 1, 2, \ldots, n)$ the limiting absolute errors of the arguments x_i and by Δ_u the limiting error of the function u, we get, for small Δx_i,

$$\Delta_u = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \Delta_{x_i} \tag{2}$$

Dividing both sides of inequality (1) by u, we get an estimate for the relative error of the function u:

$$\delta \leqslant \sum_{i=1}^{n} \left|\frac{\frac{\partial f}{\partial x_i}}{u}\right| |\Delta x_i| = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln f(x_1, \ldots, x_n)\right| |\Delta x_i| \tag{3}$$

Hence, for the limiting relative error of the function u we can take

$$\delta_u = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln u\right| \Delta_{x_i} \tag{4}$$

Example 1. Find the limiting absolute and relative errors of the volume of a sphere $V = \frac{1}{6}\pi d^3$ if the diameter $d = 3.7$ cm $\pm$ 0.05 cm and $\pi \approx 3.14$.

Solution. Regarding π and d as variable quantities, we compute the partial derivatives

$$\frac{\partial V}{\partial \pi} = \frac{1}{6} d^3 = 8.44,$$

$$\frac{\partial V}{\partial d} = \frac{1}{2} \pi d^2 = 21.5$$

By virtue of formula (2), the limiting absolute error of the volume is

$$\Delta_V = \left|\frac{\partial V}{\partial \pi}\right| |\Delta\pi| + \left|\frac{\partial V}{\partial d}\right| |\Delta d| = 8.44 \cdot 0.0016 + $$
$$+ 21.5 \cdot 0.05 = 0.013 + 1.075 = 1.088 \text{ cm}^3 \approx 1.1 \text{ cm}^3$$

and so

$$V = \frac{1}{6}\pi d^3 \approx 27.4 \text{ cm}^3 \pm 1.1 \text{ cm}^3 \qquad (5)$$

Whence the limiting relative error of the volume is

$$\delta_V = \frac{1.088 \text{ cm}^3}{27.4 \text{ cm}^3} = 0.0397 \approx 4\%$$

Example 2. Young's modulus is determined from the deflection of a rod of rectangular cross-section by the formula

$$E = \frac{1}{4} \cdot \frac{l^3 p}{a^3 b s}$$

where l is the length of the rod, a and b are the dimensions of the cross-section, s is the bending deflection, and p is the load.

Compute the limiting relative error in a determination of Young's modulus E if $p = 20$ kgf, $\delta_p = 0.1\%$, $a = 3$ mm, $\delta_a = 1\%$, $b = 44$ mm, $\delta_b = 1\%$, $l = 50$ cm, $\delta_l = 1\%$, $s = 2.5$ cm, $\delta_s = 1\%$.

Solution. $\ln E = 3\ln l + \ln p - 3\ln a - \ln b - \ln s - \ln 4$.

Whence, replacing the increments by differentials, we get

$$\frac{\Delta E}{E} = 3\frac{\Delta l}{l} + \frac{\Delta p}{p} - 3\frac{\Delta a}{a} - \frac{\Delta b}{b} - \frac{\Delta s}{s}$$

Hence

$$\delta_E = 3\delta_l + \delta_p + 3\delta_a + \delta_b + \delta_s = 3 \cdot 0.01 + 0.001 +$$
$$+ 3 \cdot 0.01 + 0.01 + 0.01 = 0.081$$

Thus, the limiting relative error constitutes 0.081, which is roughly 8% of the quantity being measured.

Performing the numerical computations, we obtain

$$E = (2.10 \pm 0.17) \cdot 10^6 \frac{\text{kgf}}{\text{cm}^2}$$

1.17 THE INVERSE PROBLEM OF THE THEORY OF ERRORS

Also of practical importance is the inverse problem: what must the absolute errors of the arguments of a function be so that the absolute error of the function does not exceed a given magnitude?

This problem is mathematically indeterminate since the given limiting error Δ_u of a function $u = f(x_1, x_2, \ldots, x_n)$ may be ensured by establishing the limiting absolute errors Δ_{x_i} of its arguments in different ways.

The simplest solution of the inverse problem is given by the so-called *principle of equal effects*. According to this principle, it is assumed that all the partial differentials

$$\frac{\partial f}{\partial x_i}\Delta x_i \qquad (i = 1, 2, \ldots, n)$$

contribute an equal amount to the total absolute error Δ_u of the function $u = f(x_1, x_2, \ldots, x_n)$.

Suppose the magnitude of the limiting absolute error Δ_u is given. Then, on the basis of formula (2) of Sec. 1.16,

$$\Delta_u = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \Delta_{x_i} \tag{1}$$

Assuming that all terms are equal, we have

$$\left|\frac{\partial u}{\partial x_1}\right| \Delta_{x_1} = \left|\frac{\partial u}{\partial x_2}\right| \Delta_{x_2} = \ldots = \left|\frac{\partial u}{\partial x_n}\right| \Delta_{x_n} = \frac{\Delta_u}{n}$$

whence

$$\Delta_{x_i} = \frac{\Delta_u}{n\left|\frac{\partial u}{\partial x_i}\right|} \qquad (i = 1, 2, \ldots, n) \tag{2}$$

Example 1. The base of a cylinder has radius $R \approx 2$ m, the altitude of the cylinder is $H \approx 3$ m. With what absolute errors must we determine R and H so that the volume V may be computed to within 0.1 m^3?

Solution. We have $V = \pi R^2 H$ and $\Delta_V = 0.1$ m^3.

Putting $R = 2$ m, $H = 3$ m, $\pi = 3.14$, we get, approximately,

$$\frac{\partial V}{\partial \pi} = R^2 H = 12,$$

$$\frac{\partial V}{\partial R} = 2\pi R H = 37.7,$$

$$\frac{\partial V}{\partial H} = \pi R^2 = 12.6$$

From this, since $n = 3$, we have, on the basis of formula (2),

$$\Delta_\pi = \frac{0.1}{3\cdot 12} < 0.003,$$

$$\Delta_R = \frac{0.1}{3\cdot 37.7} < 0.001,$$

$$\Delta_H = \frac{0.1}{3\cdot 12.6} < 0.003$$

Example 2. It is required to find the value of the function

$$u = 6x^2(\log_{10} x - \sin 2y)$$

to two decimal places; the approximate values of x and y are 15.2° and 57°, respectively. Find the permissible absolute errors in these quantities.

Solution. Here

$$u = 6x^2(\log_{10} x - \sin 2y) = 6\,(15.2)^2(\log_{10} 15.2 - \sin 114°) = 371.9,$$

$$\frac{\partial u}{\partial x} = 12x\,(\log_{10} x - \sin 2y) + 6xM = 88.54$$

where $M = 0.43429$ is the modulus of common logarithms;

$$\frac{\partial u}{\partial y} = -\,12x^2 \cos 2y = +\,1127.7$$

For the result to be correct to two decimals, the equation $\Delta_u = 0.005$ must hold. Then, by the principle of equal effects, we have

$$\Delta_x = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial x}\right|} = \frac{0.005}{2\cdot 88.54} = 0.000028,$$

$$\Delta_y = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial y}\right|} = \frac{0.005}{2\cdot 1127.7} = 0.0000022 \text{ radian} = 0''.45$$

It often happens that when solving the inverse problem by the principle of equal effects we may find that the limiting absolute errors [found from formula (2)] of the separate independent variables turn out to be so small that it is practically impossible to attain the necessary accuracy in measuring these quantities. In such cases it is best to depart from the principle of equal effects and by reasonably diminishing the errors of one part of the variables increase the errors of the other part.

Example 3. How accurately do we have to measure the radius of a circle $R = 30.5$ cm and to how many decimals do we have to take π so that the area of the circle is found to within 0.1%?

Solution. We have $s = \pi R^2$ and $\ln s = \ln \pi + 2\ln R$, whence

$$\frac{\Delta_s}{s} = \frac{\Delta_\pi}{\pi} + \frac{2\Delta_R}{R} = 0.001$$

By the principle of equal effects, put

$$\frac{\Delta_\pi}{\pi} = 0.0005, \quad \frac{2\Delta_R}{R} = 0.0005$$

whence $\Delta_\pi \leqslant 0.0016$ and $\Delta_R \leqslant 0.00025R = 0.0076$ cm.

Thus, we have to take $\pi = 3.14$ and measure R to within thousandths of a centimetre. It is plain that such accuracy of measurement is extremely difficult, and so it is better to do as follows: take $\pi =$

$= 3.142$, whence $\frac{\Delta_\pi}{\pi} = 0.00013$; then $\frac{2\Delta_R}{R} = 0.001 - 0.00013 = 0.00087$ and $\Delta_R \leqslant 0.013$ cm. Accuracy of this degree can be achieved with relative ease.

Sometimes the limiting absolute error of all arguments $x_i\,(i = 1, 2, \ldots, n)$ is allowed to be the same. Then, putting

$$\Delta_{x_1} = \Delta_{x_2} = \ldots = \Delta_{x_n}$$

we get, from formula (1),

$$\Delta_{x_i} = \frac{\Delta_u}{\sum\limits_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right|} \qquad (i = 1, 2, \ldots, n)$$

Finally, it may be assumed that the accuracy of measurements of all arguments $x_i\,(i = 1, 2, \ldots, n)$ is the same, that is, the limiting relative errors $\delta_{x_i}\,(i = 1, 2, \ldots, n)$ of the arguments are equal:

$$\delta_{x_1} = \delta_{x_2} = \ldots = \delta_{x_n}$$

From this we obtain

$$\frac{\Delta_{x_1}}{|x_1|} = \frac{\Delta_{x_2}}{|x_2|} = \ldots = \frac{\Delta_{x_n}}{|x_n|} = k$$

where k is the common value of the ratios.

Consequently

$$\Delta_{x_i} = k\,|x_i| \qquad (i = 1, 2, \ldots, n)$$

Putting these values into formula (1), we get

$$\Delta_u = k \sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|$$

and

$$k = \frac{\Delta_u}{\sum\limits_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|}$$

And we finally have

$$\Delta_{x_i} = \frac{|x_i|\,\Delta_u}{\sum\limits_{j=1}^{n} \left| x_j \frac{\partial u}{\partial x_j} \right|} \qquad (i = 1, 2, \ldots, n)$$

Other versions can also be utilized.

The solution is similar for the second inverse problem of the theory of errors when the limiting relative error of a function is given and the limiting absolute or relative errors of the argument are sought.

Occasionally, there are conditions in the very statement of the problem that prevent one from applying the principle of equal effects.

Example 4. The sides of a rectangle are $a \approx 5$ m and $b \approx 200$ m. What is the permissible limiting absolute error in measuring these sides (the same for both sides) so that the area S of the rectangle can be determined with a limiting absolute error of $\Delta_S = 1$ m²?

Solution. Since

$$S = ab$$

it follows that

$$\Delta S \approx b\Delta a + a\Delta b$$

and

$$\Delta_S = b\Delta_a + a\Delta_b$$

By hypothesis,

$$\Delta_a = \Delta_b$$

and so

$$\Delta_a = \frac{\Delta_S}{a+b} = \frac{1}{205} \approx 0.005 \text{ m} = 5 \text{ mm}$$

1.18 ACCURACY IN THE DETERMINATION OF ARGUMENTS FROM A TABULATED FUNCTION

It is often necessary in computational work to determine an argument from the value of a tabulated function. For example, we are constantly called upon to determine a number from a table of logarithms or an angle from the tabular value of a trigonometric function, etc. Naturally, any error in the function causes an error in the determination of the argument.

Suppose we have a table of single entry for the function $y = f(x)$. If the function $f(x)$ is differentiable, then for sufficiently small values $|\Delta x|$ we have

$$|\Delta y| = |f'(x)|\,|\Delta x|$$

whence

$$|\Delta x| = \frac{|\Delta y|}{|f'(x)|} \tag{1}$$

or

$$\Delta_x = \frac{1}{|y'|}\Delta_y$$

Let us apply formula (1) to some of the more common tabulated functions.

A. LOGARITHMS

Suppose $y = \ln x$, then $y' = \frac{1}{x}$.

From this

$$\Delta_x = x\Delta_y \tag{2}$$

But if $y = \log_{10} x$, then $y' = \frac{M}{x}$ where $M = 0.43429$;

$$\Delta_x = \frac{1}{M} x\Delta_y = 2.30\, x\Delta_y \tag{2'}$$

Whence, in particular, we obtain $\delta_x = 2.30\,\Delta_y$; that is, the limiting relative error of the number in the table of common logarithms is just about equal to $2^1/_2$ times the limiting absolute error of the logarithm of this number.

B. TRIGONOMETRIC FUNCTIONS

1. If $y = \sin x \left(0 < x < \frac{\pi}{2}\right)$, then $y' = \cos x$ and hence

$$\Delta_x = \Delta_y \sec x \text{ radians} \tag{3}$$

2. For the function

$$y = \tan x \left(0 < x < \frac{\pi}{2}\right)$$

we have

$$y' = \sec^2 x$$

and

$$\Delta_x = \Delta_y \cos^2 x \text{ radians} \tag{4}$$

3. If $y = \log_{10} (\sin x) \left(0 < x < \frac{\pi}{2}\right)$, then

$$y' = M \cot x \quad \text{and} \quad \Delta_x = 2.30 \tan x\Delta_y \text{ radians} \tag{5}$$

4. Putting $y = \log_{10} (\tan x) \left(0 < x < \frac{\pi}{2}\right)$, we have

$$y' = \frac{2M}{\sin 2x} \quad \text{and} \quad \Delta_x = 1.15 \sin 2x\Delta_y \text{ radians} \tag{6}$$

Since it is obvious that $\frac{\sin 2x}{2} < \tan x$ for $0 < x < \frac{\pi}{2}$, it follows from formulas (5) and (6) that the angle x is more accurately determined from a table of log tangents than from a table of log sines.

C. EXPONENTIAL FUNCTIONS

If $y=e^x$, then $y'=e^x$ and

$$\Delta_x = \frac{\Delta_y}{e^x} \tag{7}$$

or

$$\Delta_x = \frac{\Delta_y}{y}$$

Example 1. To what accuracy can we determine the number $x \approx 5000$, using a four-place table of common logarithms?

Solution. From (2′) we get

$$\Delta_x = 2.30 \cdot 5000 \cdot \frac{1}{2} \cdot 10^{-4} \approx 0.6$$

which means the number x is correct to roughly four digits.

Example 2. Find the error in a determination of the angle $x \approx 60°$ from:

(a) a five-place table of log sines,
(b) a five-place table of log tangents.

Solution. For the first case, we have, from formula (5),

$$\Delta_x = 2.30 \cdot \sqrt{3} \cdot \frac{1}{2} \cdot 10^{-5} \text{ radian} = 0.00002 \text{ radian} \approx 4''$$

In the second case, using (6), we get

$$\Delta_x = 1.15 \cdot \sqrt{3} \cdot \frac{1}{2} \cdot 10^{-5} \text{ radian} \approx 0.000005 \text{ radian} \approx 1''$$

which is one fourth the preceding error.

1.19 THE METHOD OF BOUNDS

The commonly used error estimation of a function [Sec. 1.16, formula (2)] is approximate, since this estimate is based on neglecting the products of errors. In certain cases we want to know the **exact bounds** of the required value of the function if the range of its arguments is known. The simplest way to attain this is via the *method of double computation*, also called the *method of bounds.*

Let

$$u = f(x_1, x_2, \ldots, x_n)$$

be a continuously differentiable function monotonic with respect to each argument x_i $(i = 1, 2, \ldots, n)$. For this purpose it suffices to assume that the derivatives $\frac{\partial f}{\partial x_i}$ $(i = 1, 2, \ldots, n)$ retain the same sign throughout the range ω of the arguments. Suppose that

$$\underline{x}_i < x_i < \bar{x}_i \qquad (i = 1,\ 2,\ \ldots,\ n) \tag{1}$$

and the parallelepiped (1) lies entirely in the domain ω.

Assume that $\tilde{x}_i = \underline{x}_i$, $\hat{x}_i = \bar{x}_i$ if the function f is an increasing function with respect to the variable x_i, and $\tilde{x}_i = \bar{x}_i$, $\hat{x}_i = \underline{x}_i$ if the function f is a decreasing function with respect to the variable x_i.

Then obviously

$$\underline{u} < u < \bar{u} \tag{2}$$

where

$$\underline{u} = f(\tilde{x}_1,\ \tilde{x}_2,\ \ldots,\ \tilde{x}_n)$$

and

$$\bar{u} = f(\hat{x}_1,\ \hat{x}_2,\ \ldots,\ \hat{x}_n)$$

Note that the variables $\tilde{x}_i$ $(i = 1,\ 2,\ \ldots,\ n)$ and the result of the operations of f on them can only be rounded in the sense of rounding down the quantity $\underline{u}$, and the variables $\hat{x}_i$ $(i = 1,\ 2,\ \ldots,\ n)$ and the result of the operations of f on them, only in the sense of rounding up the quantity $\bar{u}$. Strict fulfillment of inequality (2) will then be guaranteed. In the particular case when the function f is monotonically increasing with respect to each argument x_i $(i = 1,\ 2,\ \ldots,\ n)$, we simply have

$$f(\underline{x}_1,\ \underline{x}_2,\ \ldots,\ \underline{x}_n) < u < f(\bar{x}_1,\ \bar{x}_2,\ \ldots,\ \bar{x}_n) \tag{3}$$

Example. An aluminium cylinder with base diameter $d = 2\text{ cm} \pm 0.01$ cm and altitude $h = 11$ cm $\pm$ 0.02 cm weighs $p = 93.4$ gf $\pm$ 0.001 gf. Determine the specific weight γ of aluminium and estimate its limiting absolute error.

Solution. The volume of the cylinder is

$$v = \frac{\pi d^2}{4} h$$

whence

$$\gamma = \frac{p}{v} = \frac{4p}{\pi d^2 h} \tag{4}$$

From formula (4) it follows that in the range $p > 0$, $d > 0$, $h > 0$, the function γ is increasing with respect to the argument p and decreasing with respect to the arguments d and h. By hypothesis we have

$$1.99\text{ cm} \leqslant d \leqslant 2.01\text{ cm},$$
$$10.98\text{ cm} \leqslant h \leqslant 11.02\text{ cm},$$
$$93.399\text{ gf} \leqslant p \leqslant 93.401\text{ gf}$$

Besides,

$$3.14159 < \pi < 3.1416$$

And so

$$\underline{\gamma} = \frac{4\cdot 93.399}{3.1416\cdot 2.01^2\cdot 11.02} = 2.671\,\frac{\text{gf}}{\text{cm}^3}$$

(too small) and

$$\overline{\gamma} = \frac{4\cdot 93.401}{3.14159\cdot 1.99^2\cdot 10.98} = 2.735\,\frac{\text{gf}}{\text{cm}^3}$$

(too large). Taking the arithmetic mean, we get

$$\gamma = 2.703\frac{\text{gf}}{\text{cm}^3} \pm 0.027\,\frac{\text{gf}}{\text{cm}^3} \qquad (5)$$

or, after rounding,

$$\gamma = 2.70\,\frac{\text{gf}}{\text{cm}^3} \pm 0.03\,\frac{\text{gf}}{\text{cm}^3}$$

By way of comparison we give an approximate estimate of the error. Using the mean values of the arguments, we obtain

$$\gamma = \frac{4\cdot 93.4}{3.1416\cdot 2^2\cdot 11} = 2.703\,\frac{\text{gf}}{\text{cm}^3}$$

Taking logs of (4), we have

$$\ln\gamma = \ln 4 + \ln p - \ln\pi - 2\ln d - \ln h$$

whence, taking the total differential, we obtain

$$\frac{\Delta\gamma}{\gamma} = \frac{\Delta p}{p} - \frac{\Delta\pi}{\pi} - \frac{2\Delta d}{d} - \frac{\Delta h}{h}$$

Consequently

$$\delta_\gamma = \delta_p + \delta_\pi + 2\delta_d + \delta_h = \frac{0.001}{93.4} + \frac{0.00001}{3.1416} + \frac{2\cdot 0.01}{2} + \frac{0.02}{11} =$$
$$= 1.07\cdot 10^{-5} + 3.18\cdot 10^{-6} + 10^{-2} + 1.82\cdot 10^{-3} = 1.183\cdot 10^{-2}$$

Then we find

$$\Delta_\gamma = \delta_\gamma\cdot\gamma = 1.183\cdot 10^{-2}\cdot 2.703 = 3.2\cdot 10^{-2}\,\frac{\text{gf}}{\text{cm}^3}$$

Thus we have, approximately,

$$\gamma = 2.703\,\frac{\text{gf}}{\text{cm}^3} \pm 0.032\,\frac{\text{gf}}{\text{cm}^3}$$

which is very close to the exact evaluation (5).

*1.20 THE NOTION OF A PROBABILITY ERROR ESTIMATE

Suppose we have a sum of n terms:

$$u = x_1 + x_2 + \ldots + x_n$$

Then, as we know, the limiting absolute error of the sum is

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n} \tag{1}$$

Whence, when the limiting absolute errors of the terms are the same,

$$\Delta_{x_1} = \Delta_{x_2} = \ldots = \Delta_{x_n} = \Delta$$

we have

$$\Delta_u = n\Delta \tag{1'}$$

Formula (1) yields the **maximum** possible value of the absolute error of the sum. This limiting error is only attained when the errors of all the terms: (1) are the largest possible, and (2) have the same signs. Given a large number of terms, such an unfavourable coincidence has only a remote possibility. Actually, the errors in the separate terms are, as a rule, of opposite sign and consequently partially neutralize one another. For this reason, besides the theoretical limiting error Δ_u of a sum, we introduce the *practical limiting error* Δ_u^*, which occurs with a certain measure of certainty.

We confine ourselves to an elementary case. Let the absolute errors Δx_i ($i = 1, 2, \ldots, n$) of the terms of sum (1) be independent and obey the normal law with the same measure of accuracy. Assume that the absolute errors of the terms do not exceed the number Δ with a probability exceeding the number γ; that is,

$$P(|\Delta x_i| \leqslant \Delta) > \gamma$$

Given this condition, probability theory proves that the absolute error of the sum u will satisfy the inequality $|\Delta u| \leqslant \Delta\sqrt{n}$, where n is the number of terms, with the same measure of certainty.

Thus, for the limiting absolute error of a sum we can take the number

$$\Delta_u^* = \Delta\sqrt{n} \tag{2}$$

For example, in adding 100 numbers with absolute error 0.1, we get a theoretical limiting error of the sum Δ_u equal to $0.1 \cdot 100 = 10$. Actually, we may expect that this error does not exceed $0.1 \cdot 10 = 1$.

As a particular instance, consider the arithmetic mean of n numbers:

$$\xi = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

According to strict theory, the limiting absolute error is

$$\Delta_\xi = \frac{1}{n} \cdot n\Delta = \Delta$$

whereas with great certainty we can assert that in actuality

$$\Delta_{\xi}^{*}=\frac{\Delta\sqrt{n}}{n}=\frac{\Delta}{\sqrt{n}}$$

which is to say that we may be practically certain that *the arithmetic mean of approximate numbers is more accurate than the numbers themselves*, and

$$\Delta_{\xi}^{*}\longrightarrow 0 \text{ as } n\longrightarrow\infty$$

Similarly, for the case of multiplying n factors with the same limiting relative error δ, we can prove that the practical limiting relative error of the product is given by the formula

$$\delta_{u}^{*}=\delta\sqrt{n} \tag{3}$$

REFERENCES FOR CHAPTER 1

[1] *A. N. Krylov, Lectures on Approximate Computations,* 1933, Chapter I (in Russian).
[2] *D. A. Ventsel, E. S. Ventsel, Elements of the Theory of Approximate Computations,* 1949, Chapter I (in Russian).
[3] *James B. Scarborough, Numerical Mathematical Analysis,* 1955, Chapter I.
[4] *Ya. S. Bezikovich, Approximate Computations,* 1949, Chapters I and II (in Russian).
[5] *G. M. Fikhtengolts, Mathematics for Engineers,* 1933, Part 1, Chapter I (in Russian).
[6] V. M. Bradis, Oral and Written Computing. Computational Aids. *Encyclopedia of Elementary Mathematics,* Book 1, 1951 (in Russian).

Chapter 2

SOME FACTS FROM THE THEORY OF CONTINUED FRACTIONS

2.1 THE DEFINITION OF A CONTINUED FRACTION

An expression of the form

$$a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cfrac{b_3}{a_3+\dots}}}=\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \frac{b_3}{a_3},\ \dots\right] \tag{1}$$

is called a *continued fraction*. The following abbreviated notation is also used for the continued fraction (1):

$$a_0+\frac{b_1|}{|a_1}+\frac{b_2|}{|a_2}+\dots$$

In the general case, the *elements* of a continued fraction a_0, a_k, $b_k\,(k=1,\ 2,\ \dots)$ are real or complex numbers, or functions of one or more variables. The fractions $a_0=\frac{a_0}{1}$, $\frac{b_k}{a_k}\,(k=1,2,\dots)$ are called *components* of the continued fraction (1) (the zeroth, first, etc.), and the numbers of functions a_k and $b_k\,(k\geqslant 1)$ are called the *terms* of the kth component (partial denominators or numerators). We will assume that $a_k\neq 0$. Note that in the abbreviated notation (1) the components $\frac{b_k}{a_k}$ cannot be reduced.

If the continued fraction (1) contains a finite number of components (say n, not counting the zeroth one), it is called a *finite* or *n-component* continued fraction and is symbolized compactly as

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \dots,\ \frac{b_n}{a_n}\right]=\left[a_0;\ \frac{b_k}{a_k}\right]_1^n \tag{2}$$

A finite continued fraction is identified with the corresponding common fraction obtained by performing the indicated operations. A continued fraction (1) having an infinity of components is termed an *infinite* continued fraction and is denoted as

$$\left[a_0;\ \frac{b_k}{a_k}\right]_1^{\infty} \tag{3}$$

The continued fraction

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdots}}=\left[a_0;\ \frac{1}{a_1},\ \frac{1}{a_2},\ \ldots\right] \tag{4}$$

where all partial numerators are equal to 1 is termed a *simple* (or *standard*) *continued fraction. The denominators of the components* are called *partial quotients.* Note that in the theory of numbers, partial quotients are usually natural numbers (positive integers).

2.2 CONVERTING A CONTINUED FRACTION TO A SIMPLE FRACTION AND VICE VERSA

Any finite continued fraction may be converted to a simple fraction. To do this, simply perform all the operations indicated by the continued fraction.

Example 1. Change the continued fraction

$$\left[3;\ \frac{1}{3},\ \frac{1}{1},\ \frac{1}{4}\right]=3+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{4}}}$$

to a simple fraction.

Solution. Performing the indicated operations in succession, we get

$$(1)\ 1+\frac{1}{4}=\frac{5}{4},\qquad (4)\ 1:\frac{19}{5}=\frac{5}{19},$$

$$(2)\ 1:\frac{5}{4}=\frac{4}{5},\qquad (5)\ 3+\frac{5}{19}=\frac{62}{19}$$

$$(3)\ 3+\frac{4}{5}=\frac{19}{5},$$

Hence

$$\left[3;\ \frac{1}{3},\ \frac{1}{1},\ \frac{1}{4}\right]=3\frac{5}{19}$$

Conversely, any positive rational number may be converted to a continued fraction with natural elements. Suppose, for example, we are given the fraction $\frac{p}{q}$. Eliminating the integral part a_0, we have

$$\frac{p}{q}=a_0+\frac{r_0}{q}$$

where r_0 is the remainder $\left(\text{if } \frac{p}{q} \text{ is a proper fraction then } a_0=0 \text{ and } r_0=p\right)$.

Dividing the numerator and denominator of the fraction $\frac{r_0}{q}$ by r_0, we have

$$\frac{r_0}{q}=\frac{1}{q:r_0}=\frac{1}{a_1+\frac{r_1}{r_0}}$$

where a_1 is an integral quotient and r_1 is the remainder left from dividing q by r_0.

Dividing the numerator and the denominator of the fraction $\frac{r_1}{r_0}$ by r_1, we obtain

$$\frac{r_1}{r_0}=\frac{1}{r_0:r_1}=\frac{1}{a_2+\frac{r_2}{r_1}}$$

where a_2 is an integral quotient and r_2 is the remainder left from dividing r_0 by r_1. The process may be continued in similar fashion.

Since $q>r_0>r_1>r_2>r_3>\dots$ and $r_i\,(i=0, 1, 2, \dots)$ are positive integers, we will finally have $r_n=0$, or

$$\frac{r_{n-1}}{r_{n-2}}=\frac{1}{a_n+0}$$

Substituting the expressions of the fractions $\frac{r_i}{r_{i-1}}$, we get

$$\frac{p}{q}=a_0+\frac{r_0}{q}=a_0+\frac{1}{a_1+\frac{r_1}{r_0}}=a_0+\frac{1}{a_1+\cfrac{1}{a_2+\frac{r_2}{r_1}}}=$$

$$=a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdot_{\cdot_{\cdot}}+\frac{1}{a_n}}}$$

Example 2. Convert $\frac{62}{19}$ into a continued fraction.

Solution. We have, successively,

$$\frac{62}{19}=3+\frac{5}{19}=3+\frac{1}{\frac{19}{5}}=3+\frac{1}{3+\frac{4}{5}}=3+\cfrac{1}{3+\cfrac{1}{\frac{5}{4}}}=3+\cfrac{1}{3+\cfrac{1}{1+\frac{1}{4}}}$$

Thus, $\frac{62}{19}=\left[3; \frac{1}{3}, \frac{1}{1}, \frac{1}{4}\right]$.

General continued fractions are transformed analogously.

Example 3. Convert the continued fraction

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=1-\cfrac{x^2}{1-\cfrac{x^2}{3-\cfrac{x^2}{5}}}$$

to a simple one.

Solution. We have

(1) $$1-\cfrac{x^2}{3-\cfrac{x^2}{5}}=1-\frac{5x^2}{15-x^2}=\frac{15-6x^2}{15-x^2},$$

(2) $$1-\cfrac{x^2}{\cfrac{15-6x^2}{15-x^2}}=1-\frac{15x^2-x^4}{15-6x^2}=\frac{15-21x^2+x^4}{15-6x^2}$$

and so

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=\frac{15-21x^2+x^4}{15-6x^2}$$

2.3 CONVERGENTS

Suppose we have a terminating or nonterminating continued fraction

$$\left[a_0;\ \frac{b_k}{a_k}\right]_1^n \tag{1}$$

The simple fraction

$$\frac{P_k}{Q_k}\equiv\left[a_0;\ \frac{b_1}{a_1},\ \ldots,\ \frac{b_k}{a_k}\right]$$

($k=1, 2, \ldots$), where $k\leqslant n$, is called the kth *convergent* of the continued fraction (1). Following Euler, we usually set

$$\frac{P_0}{Q_0}=\frac{a_0}{1},\qquad \frac{P_{-1}}{Q_{-1}}=\frac{1}{0}$$

and for definiteness we assume that

$$P_0=a_0,\qquad Q_0=1 \tag{2}$$

and

$$P_{-1}=1,\qquad Q_{-1}=0 \tag{2'}$$

When using a digital computer, the convergents

$$\frac{P_n}{Q_n}=a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cdot_{\displaystyle\cdot_{\displaystyle\cdot+\frac{b_n}{a_n}}}}}$$

are conveniently found with the aid of *Horner's scheme* (see Chapter 3) *for division:*

$$\begin{array}{ll} c_1=\frac{b_n}{a_n}, & d_1=a_{n-1}+c_1, \\ c_2=\frac{b_{n-1}}{d_1}, & d_2=a_{n-2}+c_2, \\ \dots\dots\dots & \dots\dots\dots \\ c_k=\frac{b_{n-k+1}}{d_{k-1}}, & d_k=a_{n-k}+c_k, \\ \dots\dots\dots & \dots\dots\dots \\ c_n=\frac{b_1}{d_{n-1}}, & d_n=a_0+c_n=\frac{P_n}{Q_n} \end{array}$$

The indicated sequence of operations can readily be programmed.

Theorem 1. (Law of formation of convergents). *The numbers* P_k, Q_k $(k=-1, 0, 1, 2, \dots)$, *determined from the relations*

$$P_k=a_kP_{k-1}+b_kP_{k-2}, \tag{3}$$

$$Q_k=a_kQ_{k-1}+b_kQ_{k-2} \tag{3'}$$

where

$$P_{-1}=1, \quad Q_{-1}=0, \quad P_0=a_0, \quad Q_0=1 \tag{4}$$

are, respectively, the numerators and denominators of the convergents $\frac{P_k}{Q_k}$ *of the continued fraction* (1). We shall call such convergents *canonical*.

Proof. Let R_k $(k=1, 2, \dots)$ be the successive convergents of the continued fraction (1). It is required to prove that

$$R_k=\frac{P_k}{Q_k} \qquad (k=1, 2, \dots)$$

We carry out the proof by the method of mathematical induction. When $k=1$ we have, for the convergent R_1,

$$R_1=a_0+\frac{b_1}{a_1}=\frac{a_0a_1+b_1}{a_1}$$

On the other hand, from relations (3) and (3′), we get, taking into consideration (4),

$$P_1=a_1a_0+b_1,$$
$$Q_1=a_1\cdot 1+b_1\cdot 0=a_1$$

Hence, $R_1=\frac{P_1}{Q_1}$ and for $k=1$ the assertion of the theorem holds.

Now let the theorem be true for all natural numbers not exceeding k. We will show that the theorem also holds true for the natural number $k+1$. From (3) and (3′) we obtain

$$P_{k+1}=a_{k+1}P_k+b_{k+1}P_{k-1},$$
$$Q_{k+1}=a_{k+1}Q_k+b_{k+1}Q_{k-1}$$

By the induction hypothesis we have

$$R_k=\frac{P_k}{Q_k}=\frac{a_kP_{k-1}+b_kP_{k-2}}{a_kQ_{k-1}+b_kQ_{k-2}}.$$

By the law of formation of continued fraction (1), the convergent R_{k+1} is obtained from the convergent R_k by replacing the term a_k by the sum $a_k+\frac{b_{k+1}}{a_{k+1}}$. Therefore

$$R_{k+1}=\frac{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)P_{k-1}+b_kP_{k-2}}{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)Q_{k-1}+b_kQ_{k-2}}=$$

$$=\frac{a_{k+1}(a_kP_{k-1}+b_kP_{k-2})+b_{k+1}P_{k-1}}{a_{k+1}(a_kQ_{k-1}+b_kQ_{k-2})+b_{k+1}Q_{k-1}}=\frac{a_{k+1}P_k+b_{k+1}P_{k-1}}{a_{k+1}Q_k+b_{k+1}Q_{k-1}}=\frac{P_{k+1}}{Q_{k+1}}$$

which completes the proof.

Note. Since the terms of the convergents are not defined uniquely, one cannot, in the general case, assert that the numerators and denominators of convergents of noncanonical type satisfy the equations (3) and (3′). In the sequel we assume that the convergents under consideration are canonical.

Corollary. For the simple continued fraction

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+.}}$$

the numerators and denominators of its convergents $\frac{p_k}{q_k}$ $(k=1, 2, \ldots)$ can be determined from the relations

$$\left.\begin{aligned}p_k&=a_kp_{k-1}+p_{k-2},\\ q_k&=a_kq_{k-1}+q_{k-2}\end{aligned}\right\}\qquad(3'')$$

where we put $p_0=a_0$, $p_{-1}=1$, and $q_0=1$, $q_{-1}=0$.

Note. The following scheme is convenient for finding the terms of successive convergents from formulas (3) and (3′).

k	-1	0	1	2	3	...
b_k		1	b_1	b_2	b_3	...
a_k		a_0	a_1	a_2	a_3	...
P_k	1	a_0	P_1	P_2	P_3	...
Q_k	0	1	Q_1	Q_2	Q_3	...

In this scheme, the row b_k is omitted for a simple continued fraction where $b_k = 1$ ($k = 1, 2, \ldots$) and the formulas (3″) hold.

Example 1. Compute all the convergents of the continued fraction

$$\frac{163}{59} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{4 + \cfrac{1}{1 + \frac{1}{2}}}}}$$

Solution. Using the scheme, we obtain

a_k		2	1	3	4	1	2
p_k	$p_{-1} = 1$	2	3	11	47	58	163
q_k	$q_{-1} = 0$	1	1	4	17	21	59

Thus

$$\frac{p_0}{q_0} = \frac{2}{1}, \quad \frac{p_1}{q_1} = \frac{3}{1}, \quad \frac{p_2}{q_2} = \frac{11}{4},$$

$$\frac{p_3}{q_3} = \frac{47}{17}, \quad \frac{p_4}{q_4} = \frac{58}{21}, \quad \frac{p_5}{q_5} = \frac{163}{59}$$

Example 2. Find all the convergents of the general continued fraction

$$\left[0; \frac{1}{2}, \frac{3}{4}, \frac{5}{8}, \frac{7}{16}\right]$$

Solution. Using the foregoing scheme, we have

k	-1	0	1	2	3	4
b_k		1	1	3	5	7
a_k		0	2	4	8	16
P_k	1	0	1	4	37	620
Q_k	0	1	2	11	98	1645

Whence

$$\frac{P_0}{Q_0}=\frac{0}{1}, \quad \frac{P_1}{Q_1}=\frac{1}{2}, \quad \frac{P_2}{Q_2}=\frac{4}{11}, \quad \frac{P_3}{Q_3}=\frac{37}{98}, \quad \frac{P_4}{Q_4}=\frac{620}{1645}$$

Theorem 2. *The formula*

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k} \qquad (k\geqslant 1) \tag{4'}$$

holds true for two successive convergents $\frac{P_{k-1}}{Q_{k-1}}$ *and* $\frac{P_k}{Q_k}$ *of the continued fraction* (1).

Proof. We have

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=\frac{\Delta_k}{Q_{k-1}Q_k} \tag{5}$$

where

$$\Delta_k=\begin{vmatrix} P_k & P_{k-1} \\ Q_k & Q_{k-1} \end{vmatrix}$$

Using relations (3) and (3′), we get (by virtue of familiar determinantal properties)

$$\Delta_k=\begin{vmatrix} a_kP_{k-1}+b_kP_{k-2} & P_{k-1} \\ a_kQ_{k-1}+b_kQ_{k-2} & Q_{k-1} \end{vmatrix}=b_k\begin{vmatrix} P_{k-2} & P_{k-1} \\ Q_{k-2} & Q_{k-1} \end{vmatrix}=-b_k\Delta_{k-1}$$

From this we successively obtain

$$\Delta_k=(-b_k)(-b_{k-1})\ldots(-b_1)\,\Delta_0=(-1)^k\,b_1b_2\ldots b_k\Delta_0$$

where

$$\Delta_0=\begin{vmatrix} P_0 & P_{-1} \\ Q_0 & Q_{-1} \end{vmatrix}=\begin{vmatrix} a_0 & 1 \\ 1 & 0 \end{vmatrix}=-1$$

Thus

$$\Delta_k=(-1)^{k-1}b_1b_2\ldots b_k$$

and, consequently, on the basis of formula (5) we conclude that

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}$$

Corollary 1. If $\frac{P_{k-1}}{Q_{k-1}}$ and $\frac{P_k}{Q_k}$ $(k\geqslant 1)$ are two successive convergents of the continued fraction (1), then

$$\Delta_k=P_kQ_{k-1}-P_{k-1}Q_k=(-1)^{k-1}b_1b_2\ldots b_k$$

Corollary 2. For two successive convergents $\frac{p_{k-1}}{q_{k-1}}$, $\frac{p_k}{q_k}$ $(k\geqslant 1)$ of a simple continued fraction, the following equation holds true:

$$\frac{p_k}{q_k}-\frac{p_{k-1}}{q_{k-1}}=\frac{(-1)^{k-1}}{q_{k-1}q_k} \tag{4''}$$

Theorem 3. *For two successive convergents of equal parity* $\frac{P_{k-2}}{Q_{k-2}}$ and $\frac{P_k}{Q_k}$ $(k\geqslant 2)$ *of the continued fraction* (1), *the relation*

$$\frac{P_k}{Q_k}-\frac{P_{k-2}}{Q_{k-2}}=(-1)^k\frac{b_1b_2\ldots b_{k-1}a_k}{Q_{k-2}Q_k} \tag{6}$$

holds true.

Proof. We have

$$\frac{P_k}{Q_k}-\frac{P_{k-2}}{Q_{k-2}}=\frac{D_k}{Q_{k-2}Q_k}, \tag{7}$$

where

$$D_k=\begin{vmatrix}P_k & P_{k-2}\\ Q_k & Q_{k-2}\end{vmatrix}$$

whence, on the basis of the law of formation of convergents and on the basis of elementary properties of determinants, we obtain

$$D_k=\begin{vmatrix}a_kP_{k-1}+b_kP_{k-2} & P_{k-2}\\ a_kQ_{k-1}+b_kQ_{k-2} & Q_{k-2}\end{vmatrix}=a_k\begin{vmatrix}P_{k-1} & P_{k-2}\\ Q_{k-1} & Q_{k-2}\end{vmatrix}=a_k\Delta_{k-1}$$

where Δ_k is the determinant considered in Theorem 2. By the corollary to Theorem 1, we have

$$\Delta_{k-1}=(-1)^k\,b_1b_2\ldots b_{k-1}$$

whence

$$D_k=(-1)^k\,b_1b_2\ldots b_{k-1}a_k$$

Consequently we obtain formula (6) by using relation (7).

Corollary. If $\frac{p_{k-2}}{q_{k-2}}$ and $\frac{p_k}{q_k}$ are two successive convergents of the same parity of the simple continued fraction

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdot\atop{\cdot\atop\cdot}}}$$

then we have the relation

$$\frac{p_k}{q_k}-\frac{p_{k-2}}{q_{k-2}}=(-1)^k\frac{a_k}{q_{k-2}q_k} \tag{6'}$$

Theorem 4. *If all the elements of a finite continued fraction are positive, then its convergents of even order form a monotonic increasing sequence and the convergents of odd order form a monotonic decreasing sequence. Each convergent of even order is less than any convergent of odd order. The number α itself, which is expressed by the continued fraction, lies between two successive convergents.*

Proof. Suppose we have the continued fraction

$$\alpha=\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n}\right] \tag{8}$$

with positive elements a_k and b_k and suppose $\frac{P_k}{Q_k}$ $(k=0, 1, \ldots, n)$ are its successive canonical convergents. Then obviously $P_k>0$ and $Q_k>0$.

We consider two cases.

1. Let $k=2m$ be an even number. Then from (6), taking into consideration that $a_k>0$ and $b_i>0$ $(i=1, \ldots, k)$, we have

$$\frac{P_{2m}}{Q_{2m}}-\frac{P_{2m-2}}{Q_{2m-2}}>0$$

Consequently

$$\frac{P_{2m-2}}{Q_{2m-2}}<\frac{P_{2m}}{Q_{2m}} \quad (m=1,\ 2,\ \ldots)$$

or

$$\frac{P_0}{Q_0}<\frac{P_2}{Q_2}<\frac{P_4}{Q_4}<\ldots \tag{9}$$

2. Let $k=2m+1$ be odd. Then $k-1$ will be even, and from the same relation (6) we get

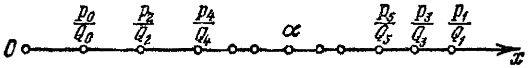
$$\frac{P_{2m-1}}{Q_{2m-1}}>\frac{P_{2m+1}}{Q_{2m+1}}$$

or

$$\frac{P_1}{Q_1}>\frac{P_3}{Q_3}>\frac{P_5}{Q_5}>\ldots \tag{10}$$

We have thus proved that even convergents form a monotonic increasing sequence and odd convergents form a monotonic decreasing sequence (Fig. 1).

Fig. 1



Furthermore, if in (4′) we set $k=2m$, we get

$$\frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}} \tag{11}$$

which is to say that every convergent of odd order is greater than the adjacent convergent of even order. We conclude therefrom that any convergent of odd order is greater than any convergent of even order. Indeed, let $\frac{P_{2s-1}}{Q_{2s-1}}$ be any odd convergent. If $s \leqslant m$, then

$$\frac{P_{2s-1}}{Q_{2s-1}} \geqslant \frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}}$$

but if $s > m$, then

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2s}}{Q_{2s}} > \frac{P_{2m}}{Q_{2m}}$$

And so for any s and m we have

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2m}}{Q_{2m}} \tag{12}$$

Finally, from the law of formation of a continued fraction

$$\alpha = a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + .}}$$

we have the obvious relations

$$\alpha > \frac{P_0}{Q_0}, \quad \alpha < \frac{P_1}{Q_1}, \quad \alpha > \frac{P_2}{Q_2}, \ \ldots$$

Hence

$$\frac{P_k}{Q_k} < \alpha < \frac{P_{k+1}}{Q_{k+1}} \tag{13}$$

if k is even and

$$\frac{P_k}{Q_k} > \alpha > \frac{P_{k+1}}{Q_{k+1}} \tag{13′}$$

if k is odd. For the last convergent, we will clearly have an equality on the right in place of the strict inequalities (13) and (13′).

Corollary 1. If the elements of the continued fraction (8) are positive and $\frac{P_k}{Q_k}$ are its convergents, then the following estimate holds true:

$$\left|\alpha-\frac{P_k}{Q_k}\right| \leqslant \frac{b_1 b_2 \ldots b_{k+1}}{Q_k Q_{k+1}} \tag{14}$$

Indeed, since, by what has been proved, we have

$$\left|\alpha-\frac{P_k}{Q_k}\right| \leqslant \left|\frac{P_{k+1}}{Q_{k+1}}-\frac{P_k}{Q_k}\right|$$

it follows, on the basis of (4′), that (14) holds true.

Corollary 2. If the continued fraction α with positive elements is simple and $\frac{p_k}{q_k}$ are its successive convergents, then

$$\left|\alpha-\frac{p_k}{q_k}\right| \leqslant \frac{1}{q_k q_{k+1}}$$

Note. If all elements of a simple continued fraction are natural, it may be demonstrated [1] that the convergent $\frac{p_k}{q_k}$ is *the best approximation* to the number α, that is, that all the remaining fractions $\frac{p}{q}$ with denominator $q \leqslant q_k$ deviate from α more than the fraction $\frac{p_k}{q_k}$.

Example 3. The second last convergent of $\frac{163}{59}$ is $\frac{58}{21}$ (see Example 1). Therefore

$$\left|\frac{163}{59}-\frac{58}{21}\right| \leqslant \frac{1}{59 \cdot 21} < 0.001$$

2.4 NONTERMINATING CONTINUED FRACTIONS

Let

$$\left[a_0; \frac{b_1}{a_1}, \frac{b_2}{a_2}, \ldots\right] = a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + {}}} \tag{1}$$

be a nonterminating continued fraction. We consider a segment, that is, a terminating continued fraction:

$$\left[a_0; \frac{b_1}{a_1}, \frac{b_2}{a_2}, \ldots, \frac{b_n}{a_n}\right] = \frac{P_n}{Q_n} \qquad (n = 1, 2, 3, \ldots) \tag{2}$$

Definition. A nonterminating continued fraction (1) is called *convergent* if there exists a finite limit

$$\alpha = \lim_{n \to \infty} \frac{P_n}{Q_n} \tag{3}$$

and the number α is taken as the value of the fraction. But if no limit (3) exists, then the continued fraction (1) is termed *divergent* and no numerical value is assigned to it.

By *Cauchy's test* [3] for the convergence of a sequence $\frac{P_n}{Q_n}$ ($n = 1, 2, 3, \ldots$) it is necessary and sufficient that there exist for every $\varepsilon > 0$ a number $N = N(\varepsilon)$ such that

$$\left| \frac{P_{n+m}}{Q_{n+m}} - \frac{P_n}{Q_n} \right| < \varepsilon$$

for $n > N$ and for any $m > 0$.

If $Q_k \neq 0$, then we obviously have

$$\frac{P_n}{Q_n} = \frac{P_0}{Q_0} + \sum_{k=1}^{n} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} \right) \tag{4}$$

From this

$$\lim_{n \to \infty} \frac{P_n}{Q_n} = \frac{P_0}{Q_0} + \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} \right) = \frac{P_0}{Q_0} + \sum_{k=1}^{\infty} (-1)^{k-1} \frac{b_1 b_2 \ldots b_k}{Q_{k-1} Q_k} \tag{4'}$$

That is, the convergence of continued fraction (1) is equivalent to the convergence of the series (4'). If the continued fraction (1) converges:

$$\alpha = \lim_{n \to \infty} \frac{P_n}{Q_n}$$

then, by virtue of formulas (4) and (4'), we have

$$\left| \alpha - \frac{P_n}{Q_n} \right| \leqslant \sum_{k=n+1}^{\infty} \left| \frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} \right| \leqslant \sum_{k=n+1}^{\infty} \left| \frac{b_1 b_2 \ldots b_k}{Q_{k-1} Q_k} \right|$$

Theorem 1. *If all the elements* a_k, b_k ($k = 0, 1, \ldots$) *of continued fraction* (1) *are positive, and*

$$b_k \leqslant a_k \quad \text{and} \quad a_k \geqslant d > 0 \qquad (k = 1, 2, \ldots) \tag{5}$$

then (1) *converges.*

Proof. When proving the first part of Theorem 4 of the preceding section we did not make use of the property of finiteness of a continued fraction. Therefore, repeating this proof, we establish the fact that if the elements of continued fraction (1) are positive, then its even convergents $\frac{P_{2k}}{Q_{2k}}$ ($k = 0, 1, 2, \ldots$) form a monotonic increasing sequence bounded from above (for example, by the number $\frac{P_1}{Q_1}$). Whence, by virtue of a familiar theorem, we conclude that the limit

$$\lim_{k\to\infty} \frac{P_{2k}}{Q_{2k}}=\alpha$$

exists. Similarly, under the conditions of our theorem odd convergents $\frac{P_{2k+1}}{Q_{2k+1}}$ $(k=0, 1, 2, \ldots)$ of the continued fraction (1) form a monotonic decreasing sequence bounded from below (by the number $\frac{P_0}{Q_0}$, for example). Hence there also exists

$$\lim_{k\to\infty} \frac{P_{2k+1}}{Q_{2k+1}}=\beta$$

and $\beta\geqslant\alpha$. Besides, for any $k\geqslant 0$ we have

$$\frac{P_{2k}}{Q_{2k}}<\alpha\leqslant\beta<\frac{P_{2k+1}}{Q_{2k+1}}$$

and so, using Theorem 2 of Sec. 2.3, we get

$$0\leqslant\beta-\alpha<\frac{P_{2k+1}}{Q_{2k+1}}-\frac{P_{2k}}{Q_{2k}}=\frac{b_1b_2\ldots b_{2k+1}}{Q_{2k}Q_{2k+1}}=\eta_k \tag{6}$$

We shall show that $\eta_k\to 0$ as $k\to\infty$. Indeed, on the basis of the law of formation of convergents, we have, for $k\geqslant 2$,

$$Q_k=a_kQ_{k-1}+b_kQ_{k-2}$$

and

$$Q_{k-1}=a_{k-1}Q_{k-2}+b_{k-1}Q_{k-3}$$

Whence, by virtue of condition (5) of the theorem, we conclude that

$$Q_k\geqslant b_k(Q_{k-1}+Q_{k-2})$$

and

$$Q_{k-1}\geqslant dQ_{k-2}$$

Hence

$$Q_k\geqslant b_k(1+d)\,Q_{k-2} \tag{7}$$

From the inequality (7) we successively obtain

$$Q_{2k}\geqslant b_{2k}(1+d)\,Q_{2k-2}\geqslant\ldots$$
$$\ldots\geqslant b_{2k}b_{2k-2}\ldots b_2(1+d)^k\,Q_0=b_2b_4\ldots b_{2k}(1+d)^k \tag{8}$$

and

$$Q_{2k+1}\geqslant b_{2k+1}(1+d)\,Q_{2k-1}\geqslant\ldots$$
$$\ldots\geqslant b_{2k+1}\ldots b_3(1+d)^k\,Q_1\geqslant b_1b_3\ldots b_{2k+1}(1+d)^k \tag{9}$$

since $Q_1=a_1\geqslant b_1$. Multiplying together inequalities (8) and (9), we get

$$Q_{2k}Q_{2k+1}\geqslant b_1b_2\ldots b_{2k+1}(1+d)^{2k} \tag{10}$$

and hence,

$$\eta_k=\frac{b_1b_2\ldots b_{2k+1}}{Q_{2k}Q_{2k+1}}\leqslant\frac{1}{(1+d)^{2k}}$$

Thus, $\eta_k\longrightarrow 0$ as $k\longrightarrow\infty$.

And so, passing to the limit as $k\longrightarrow\infty$ in inequality (6), we have $0\leqslant\beta-\alpha\leqslant 0$, or

$$\alpha=\beta=\lim_{n\to\infty}\frac{P_n}{Q_n}$$

and therefore the continued fraction (1) converges.

Note. The value α of the converging fraction (1) with positive elements lies between two successive convergents $\frac{P_{n-1}}{Q_{n-1}}$ and $\frac{P_n}{Q_n}$. Hence

$$\left|\alpha-\frac{P_n}{Q_n}\right|\leqslant\left|\frac{P_n}{Q_n}-\frac{P_{n-1}}{Q_{n-1}}\right|=\frac{b_1b_2\ldots b_n}{Q_{n-1}Q_n}$$

Corollary. A simple continued fraction with natural elements always converges.

The following theorem can also be proved [1].

Theorem 2. *Every positive number α may be expanded in unique fashion into a simple converging continued fraction with natural elements.* The resulting continued fraction is terminating if α is a rational number, and nonterminating if α is irrational.

Example. Expand the number $\sqrt{41}$ into a continued fraction and find its approximate value.

Solution. Since the largest integer in $\sqrt{41}$ is 6, we have

$$\sqrt{41}=6+\frac{1}{a_1} \tag{11}$$

whence

$$a_1=\frac{1}{\sqrt{41}-6}=\frac{6+\sqrt{41}}{5}$$

The largest integer in a_1 is 2, and so

$$a_1=2+\frac{1}{a_2} \tag{12}$$

whence

$$a_2=\frac{1}{a_1-2}=\frac{5}{\sqrt{41}-4}=\frac{4+\sqrt{41}}{5}=2+\frac{1}{a_3} \tag{13}$$

Similarly

$$a_3=\frac{1}{a_2-2}=\frac{5}{\sqrt{41}-6}=6+\sqrt{41}=12+\frac{1}{a_4}, \tag{14}$$

$$a_4=\frac{1}{a_3-12}=\frac{1}{\sqrt{41}-6}=\frac{6+\sqrt{41}}{5}=2+\frac{1}{a_5} \tag{15}$$

We note that $a_4=a_1$ and so the elements of the continued fraction will repeat, that is the continued fraction will be recurring. Substituting successively expressions (12), (13), (14), (15), etc. into (11), we get

$$\sqrt{41}=6+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{12+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{12+\ddots}}}}}}$$

Thus, the irrational number $\sqrt{41}$ is expressed as a nonterminating periodic continued fraction:

$$\sqrt{41}=\left(6;\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \ldots\right)$$

The convergents $\frac{p_k}{q_k}$ $(k=0,\ 1,\ 2,\ \ldots)$ are found by means of the following scheme:

a_k	—	6	2	2	12	2	2	12
p_k	$p_{-1}=1$	$p_0=6$	13	32	397	826	2049	...
q_k	$q_{-1}=0$	$q_0=1$	2	5	62	129	320	...

Confining ourselves, say, to the fifth convergent, we obtain a major (too large) approximation of $\sqrt{41}$: $\sqrt{41}=\frac{2049}{320}=6.403125$ with absolute error less than

$$\Delta=\frac{1}{320\,(2\cdot 320+129)}=\frac{1}{320\cdot 769}<5\cdot 10^{-6}$$

Theorem 3 (Pringsheim). *If for the nonterminating continued fraction*

$$\left[0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n},\ \ldots\right] \tag{16}$$

the inequalities

$$|b_n|+1 \leqslant |a_n| \qquad (n=1,\ 2,\ \ldots) \tag{17}$$

hold true, then this fraction converges, and its absolute value does not exceed unity [4].

Proof. Let $\frac{P_k}{Q_k}$ $(k=1,\ 2,\ \ldots)$ be convergents of the continued fraction (16).

Since $Q_k=a_kQ_{k-1}+b_kQ_{k-2}$ $(k=1,\ 2,\ \ldots)$ it follows that

$$|Q_k| \geqslant |a_k|\,|Q_{k-1}|-|b_k|\,|Q_{k-2}|$$

Whence, using inequality (17), we get

$$|Q_k| \geqslant (|b_k|+1)\,|Q_{k-1}|-|b_k|\,|Q_{k-2}|$$

or

$$|Q_k|-|Q_{k-1}| \geqslant |b_k|\,(|Q_{k-1}|-|Q_{k-2}|) \tag{18}$$

Applying inequality (18) successively and noting that $Q_0=1$ and $Q_{-1}=0$, we have

$$|Q_k|-|Q_{k-1}| \geqslant |b_k|\,|b_{k-1}|\ldots|b_1| \tag{19}$$

From inequality (19) it follows that $|Q_k|$ increases monotonically as k increases, and $|Q_k| \geqslant |Q_0|=1$.

The convergence of the continued fraction (16) is equivalent to the convergence of the series

$$\frac{P_0}{Q_0}+\sum_{k=1}^{\infty}\left(\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right)=\sum_{k=1}^{\infty}\frac{(-1)^{k-1}b_1b_2\ldots b_k}{Q_{k-1}Q_k} \tag{20}$$

Let us consider the series of the moduli

$$\sum_{k=1}^{\infty}\frac{|b_1|\,|b_2|\ldots|b_k|}{|Q_{k-1}|\,|Q_k|} \tag{21}$$

On the basis of inequality (19) we have

$$\sum_{k=1}^{n}\frac{|b_1|\,|b_2|\ldots|b_k|}{|Q_{k-1}|\,|Q_k|} \leqslant \sum_{k=1}^{n}\frac{|Q_k|-|Q_{k-1}|}{|Q_{k-1}|\,|Q_k|}=$$

$$=\sum_{k=1}^{n}\left(\frac{1}{|Q_{k-1}|}-\frac{1}{|Q_k|}\right)=\frac{1}{|Q_0|}-\frac{1}{|Q_n|}<\frac{1}{|Q_0|}=1 \quad (n=1,\ 2,\ \ldots)$$

Thus, the partial sums of the series (21) are bounded and hence the series converges, and also

$$\sum_{k=1}^{\infty} \frac{|b_1||b_2|\ldots|b_k|}{|Q_{k-1}||Q_k|} \leqslant 1 \tag{22}$$

But then the series (20) also converges (and converges absolutely) by virtue of the comparison test, that is to say there exists

$$\lim_{n\to\infty} \frac{P_n}{Q_n} = \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right) = \alpha$$

Besides, taking into account inequality (22), we have

$$|\alpha| \leqslant 1$$

Note 1. For the continued fraction (16) to converge, it is sufficient that inequality (17) hold for $n \geqslant m$; $Q_k \neq 0$ for $k \leqslant m$.

Note 2. Under the conditions of Theorem 3, we have the following estimate for the value of the continued fraction α:

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \sum_{k=n+1}^{\infty} \left|\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right| =$$

$$= \sum_{k=n+1}^{\infty} \frac{|b_1||b_2|\ldots|b_k|}{|Q_{k-1}||Q_k|} \leqslant \sum_{k=n+1}^{\infty} \frac{|Q_k| - |Q_{k-1}|}{|Q_{k-1}||Q_k|} =$$

$$= \sum_{k=n+1}^{\infty} \left(\frac{1}{|Q_{k-1}|} - \frac{1}{|Q_k|}\right) = \frac{1}{|Q_n|} - \lim_{k\to\infty} \frac{1}{|Q_k|}$$

In particular, if $|Q_k| \to +\infty$ as $k \to \infty$, then

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \frac{1}{|Q_n|}$$

2.5 EXPANDING FUNCTIONS INTO CONTINUED FRACTIONS

Continued fractions are a convenient way of representing and computing functions. For details see the special literature (for example [2]); we confine ourselves here to a selection of examples.

Note that if a function $f(x)$ can, by some technique, be expanded into a nonterminating continued fraction, then in the general case we must prove the convergence of the fraction and assure ourselves that the limiting value of the continued fraction is equal to the function $f(x)$.

A. EXPANDING A RATIONAL FUNCTION INTO A CONTINUED FRACTION

If

$$f(x)=\frac{c_{10}+c_{11}x+c_{12}x^2+\dots}{c_{00}+c_{01}x+c_{02}x^2+\dots}$$

then in the general case we have, after performing elementary manipulations,

$$f(x)=\frac{1}{\frac{c_{00}}{c_{10}}+\frac{c_{00}+c_{01}x+c_{02}x^2+\dots}{c_{10}+c_{11}x+c_{12}x^2+\dots}-\frac{c_{00}}{c_{10}}}=\frac{c_{10}}{c_{00}+xf_1(x)}$$

where

$$f_1(x)=\frac{c_{20}+c_{21}x+c_{22}x^2+\dots}{c_{10}+c_{11}x+c_{12}x^2+\dots}$$

and

$$c_{2k}=c_{10}c_{0,\,k+1}-c_{00}c_{1,\,k+1}\qquad (k=0,\ 1,\ \dots)$$

Similarly

$$f_1(x)=\frac{c_{20}}{c_{10}+xf_2(x)}$$

where

$$f_2(x)=\frac{c_{30}+c_{31}x+c_{32}x^2+\dots}{c_{20}+c_{21}x+c_{22}x^2+\dots}$$

and

$$c_{3k}=c_{20}c_{1,\,k+1}-c_{10}c_{2,\,k+1}\qquad (k=0,\ 1,\ \dots)$$

and so forth.

Thus

$$f(x)=\cfrac{c_{10}}{c_{00}+\cfrac{c_{20}x}{c_{10}+\cfrac{c_{30}x}{c_{20}+\dots}}}=\left[0;\ \frac{c_{10}}{c_{00}},\ \frac{c_{20}x}{c_{10}},\ \frac{c_{30}x}{c_{20}},\ \dots,\ \frac{c_{n0}x}{c_{n-1,\,0}}\right]\qquad(1)$$

and it is easy to see that the continued fraction (1) is a terminating fraction.

The coefficients c_{jk} of the expansions are conveniently computed successively by the formula

$$c_{jk}=-\begin{vmatrix} c_{j-2,\,0} & c_{j-2,\,k+1} \\ c_{j-1,\,0} & c_{j-1,\,k+1} \end{vmatrix}$$

where $j\geqslant 2$.

Note that in some cases the coefficients c_{jk} may prove equal to zero. Then appropriate changes must be introduced in the expansion (1) [2].

Example 1. Expand the function

$$f(x)=\frac{1-x}{1-5x+6x^2}$$

into a continued fraction.

Solution. Write down the coefficients c_{jk} in the following scheme:

$j \backslash k$	0	1	2
0	1	-5	6
1	1	-1	0
2	-4	6	0
3	-2	0	0
4	-12	0	0

Thus

$$\frac{1-x}{1-5x+6x^2}=\left[0;\ \frac{1}{1},\ \frac{-4x}{1},\ \frac{-2x}{-4},\ \frac{-12x}{-2}\right]=\cfrac{1}{1-\cfrac{4x}{1-\cfrac{2x}{-4+6x}}}\,.$$

B. EXPANDING e^x INTO A CONTINUED FRACTION

For e^x Euler obtained the expansion

$$e^x=\left[0;\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ \ldots,\ \frac{x^2}{4n+2},\ \ldots\right] \tag{2}$$

which converges for any value of x, real or complex [2].

From this we obtain the convergents

$$\frac{P_1}{Q_1}=\frac{1}{1},$$
$$\frac{P_2}{Q_2}=\frac{2+x}{2-x},$$
$$\frac{P_3}{Q_3}=\frac{12+6x+x^2}{12-6x+x^2},$$
$$\frac{P_4}{Q_4}=\frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3}$$

and so forth.

In particular, setting $x=1$ and confining ourselves to the fourth convergent, we have

$$e\approx\frac{193}{71}=2.7183\ \ldots$$

To obtain the same accuracy in the Maclaurin expansion

$$e=2+\frac{1}{2!}+\frac{1}{3!}+\dots$$

we need at least eight terms.

C. EXPANDING tan x INTO A CONTINUED FRACTION

For $\tan x$, Lambert obtained the expansion

$$\tan x=\left[0;\ \frac{x}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5},\ \dots,\ \frac{-x^2}{2n+1},\ \dots\right] \tag{3}$$

which converges at all points of continuity of the function.

Example 2. Find tan 1 approximately.

Solution. Setting $x=1$ in (3), we have

$$\tan 1=\left[0;\ \frac{1}{1},\ \frac{-1}{3},\ \frac{-1}{5},\ \dots\right]$$

On the basis of formula (3) of Sec. 2.3 let us set up the following scheme for the terms of convergents:

k	-1	0	1	2	3	4
b_k		1	1	-1	-1	1
a_k		0	1	3	5	7
P_k	1	0	1	3	14	95
Q_k	0	1	1	2	9	61

Confining ourselves to the fourth convergent, we have

$$\tan 1\approx\frac{95}{61}=1.557377$$

(tables give $\tan 1=1.557396$).

REFERENCES FOR CHAPTER 2

[1] *A. Ya. Khinchin, Continued Fractions*, 1949, Chapter I (in Russian).
[2] *A. N. Khovansky, Application of Continued Fractions and Their Generalizations to Problems of Approximate Analysis*, 1956, Chapters I and II (in Russian).
[3] *G. M. Fikhtengolts, Principles of Mathematical Analysis*, 1955, Vol. 1, Chapter III (in Russian).
[4] *O. Perron, Die Lehre von den Kettenbrüchen*, 1929, Chapter VII.

Chapter 3

COMPUTING THE VALUES OF FUNCTIONS

When using computing machines to evaluate functions given by formulas, the form in which the formula is written is not at all immaterial. Expressions which are mathematically equivalent are not of the same status as concerns approximate computations. This gives rise to a problem of practical importance, that of finding the most convenient analytic expressions for elementary functions. Computing the values of functions ordinarily reduces to a sequence of elementary arithmetic operations. Taking into account the restricted storage volume of any machine, it is desirable to split up the operations into repeating cycles. We now consider some typical techniques of computation.

3.1 COMPUTING THE VALUES OF A POLYNOMIAL. HORNER'S SCHEME

Suppose we have a polynomial of degree n,

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \tag{1}$$

with real coefficients a_k $(k = 0, 1, \ldots, n)$. Suppose it is required to find the value of this polynomial for $x = \xi$:

$$P(\xi) = a_0\xi^n + a_1\xi^{n-1} + \ldots + a_{n-1}\xi + a_n \tag{2}$$

The computation of $P(\xi)$ is most conveniently performed as follows. Represent formula (2) as

$$P(\xi) = (\ldots(((a_0\xi + a_1)\,\xi + a_2)\,\xi + a_3)\,\xi + \ldots + a_{n-1})\,\xi + a_n)$$

We then successively compute the numbers

$$\left.\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + b_0\xi, \\ b_2 &= a_2 + b_1\xi, \\ b_3 &= a_3 + b_2\xi, \\ &\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ b_n &= a_n + b_{n-1}\xi \end{aligned}\right\} \tag{3}$$

and find $b_n = P(\xi)$.

We will show that the numbers $b_0=a_0$, b_1, ..., b_{n-1} are coefficients of the polynomial $Q(x)$ obtained as the quotient on the division of the given polynomial $P(x)$ by the binomial $x-\xi$. Indeed, let

$$Q(x)=\beta_0x^{n-1}+\beta_1x^{n-2}+\dots+\beta_{n-1} \tag{4}$$

and

$$P(x)=Q(x)(x-\xi)+\beta_n \tag{5}$$

On the basis of the remainder theorem the remainder is $\beta_n=P(\xi)$. From formulas (4) and (5) we get

$$P(x)=(\beta_0x^{n-1}+\beta_1x^{n-2}+\dots+\beta_{n-1})(x-\xi)+\beta_n$$

or, removing brackets and collecting terms,

$$\begin{aligned}P(x)=\beta_0x^n+(\beta_1-\beta_0\xi)x^{n-1}+(\beta_2-\beta_1\xi)x^{n-2}+\dots+\\+(\beta_{n-1}-\beta_{n-2}\xi)x+(\beta_n-\beta_{n-1}\xi)\end{aligned}$$

Comparing coefficients of identical powers of the variable x in the right and left members of this equation, we find

$$\begin{aligned}\beta_0&=a_0,\\ \beta_1-\beta_0\xi&=a_1,\\ \beta_2-\beta_1\xi&=a_2,\\ &\dots\dots\dots\\ \beta_{n-1}-\beta_{n-2}\xi&=a_{n-1},\\ \beta_n-\beta_{n-1}\xi&=a_n\end{aligned}$$

whence

$$\begin{aligned}\beta_0&=a_0=b_0,\\ \beta_1&=a_1+\beta_0\xi=b_1,\\ \beta_2&=a_2+\beta_1\xi=b_2,\\ &\dots\dots\dots\dots\\ \beta_{n-1}&=a_{n-1}+\beta_{n-2}\xi=b_{n-1},\\ \beta_n&=a_n+\beta_{n-1}\xi=b_n\end{aligned}$$

which completes the proof.

Thus, formulas (3) enable us, without performing the operation of division, to determine the coefficients of the quotient $Q(x)$, and also the remainder $P(\xi)$. Practical computations are carried out in accord with a scheme called *Horner's scheme:*

$$\begin{array}{ccccccc} & a_0 & a_1 & a_2 & \dots & a_n & \underline{|\xi} \\ + & & b_0\xi & b_1\xi & \dots & b_{n-1}\xi & \\ \hline & b_0 & b_1 & b_2 & \dots & \underline{b_n=P(\xi)} & \end{array}$$

Example 1. Compute the values of the polynomial

$$P(x)=3x^3+2x^2-5x+7 \quad \text{for} \quad x=3$$

Solution. Set up the Horner scheme:

$$\begin{array}{rrrrl}
+\;3 & 2 & -5 & 7 & \quad \underline{|3} \\
 & 9 & 33 & 84 & \\
\hline
3 & 11 & 28 & \underline{\underline{91=P(3)}} &
\end{array}$$

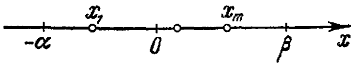
Note. Horner's scheme permits us to find the bounds of the real roots of the given polynomial $P(x)$.

Suppose that for $x=\beta$ $(\beta>0)$ all the coefficients b_i in the Horner scheme are nonnegative, and the first coefficient is positive, that is,

$$b_0=a_0>0 \quad \text{and} \quad b_i \geqslant 0 \quad (i=1, 2, \ldots, n) \tag{6}$$

We can then assert that all the real roots x_k $(k=1, 2, \ldots, m;\ m \leqslant n)$ of the polynomial $P(x)$ are not situated to the right of β, that is, $x_k \leqslant \beta$ $(k=1, 2, \ldots, m)$ (Fig. 2).

Fig. 2



Indeed, since

$$P(x)=(b_0x^{n-1}+\ldots+b_{n-1})(x-\beta)+b_n$$

it follows that for any $x>\beta$ we have, by virtue of condition (6), $P(x)>0$, which is to say that any number exceeding β is definitely not a root of the polynomial $P(x)$. We thus have an upper bound for the real roots x_k of a polynomial.

To obtain a lower bound of the roots x_k, form the polynomial

$$(-1)^n P(-x)=a_0x^n-a_1x^{n-1}+\ldots+(-1)^n a_n$$

For this new polynomial, we find a number $x=\alpha$ $(\alpha>0)$ such that all coefficients in Horner's scheme are nonnegative. Then, by the earlier reasoning, for the real roots of the polynomial $(-1)^n P(-x)$, which are obviously equal to $-x_k$ $(k=1, 2, \ldots, m)$, we have the inequality $-x_k \leqslant \alpha$.

Hence, $x_k \geqslant -\alpha$ $(k=1, 2, \ldots, m)$. We have thus obtained the lower bound $-\alpha$ of the real roots of the polynomial $P(x)$, whence it follows that all real roots of $P(x)$ lie in the interval $[-\alpha, \beta]$.

Example 2. Find the bounds of the real roots of the polynomial

$$P(x)=x^4-2x^3+3x^2+4x-1$$

Solution. Compute the value of $P(x)$ for, say, $x=2$. Using Horner's scheme, we have

$$\begin{array}{crrrrrl}
 & 1 & -2 & 3 & 4 & -1 & \underline{|2} \\
+ & & 2 & 0 & 6 & 20 & \\ \hline
 & 1 & 0 & 3 & 10 & \underline{19} &
\end{array}$$

Since all coefficients $b_i \geqslant 0$, the real roots x_k of the polynomial $P(x)$ (if they exist) satisfy the inequality $x_k < 2$. The upper bound of the real roots is found. Now let us find the lower bound. Form a new polynomial:

$$Q(x) = (-1)^4 P(-x) = x^4 + 2x^3 + 3x^2 - 4x - 1$$

Computing the value of $Q(x)$ for, say, $x=1$, we have

$$\begin{array}{crrrrrl}
 & 1 & 2 & 3 & -4 & -1 & \underline{|1} \\
+ & & 1 & 3 & 6 & 2 & \\ \hline
 & 1 & 3 & 6 & 2 & 1 &
\end{array}$$

All coefficients $b_i > 0$, hence $-x_k < 1$, that is, $x_k > -1$. And so all real roots of the given polynomial lie within the interval $[-1, 2]$.

3.2 THE GENERALIZED HORNER SCHEME

Suppose that in a given polynomial

$$P(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \tag{1}$$

it is required for some reason to make a substitution of the variable x by the formula

$$x = y + \xi \tag{2}$$

where ξ is a fixed number and y is a new variable.

Substitute expression (2) into (1) and perform the indicated operations. Collecting terms we get a new polynomial in the variable y:

$$P(y+\xi) = A_0 y^n + A_1 y^{n-1} + \ldots + A_n \tag{3}$$

Since polynomial (3) may be regarded as a Taylor series expansion in powers of y of the functions $P(y+\xi)$, the coefficients A_i $(i = 0, 1, 2, \ldots, n)$ can be computed from the formula

$$A_i = \frac{P^{(n-i)}(\xi)}{(n-i)!} \quad (i = 0, 1, \ldots, n)$$

A more convenient practical method for finding the coefficients A_i $(i = 0, 1, 2, \ldots, n)$ is by means of the Horner scheme. First put $y = 0$ in (3). We then have $A_n = P(\xi)$.

Dividing the polynomial (1) by the binomial $x - \xi$, we get

$$P(x) = (x - \xi) P_1(x) + P(\xi) \tag{4}$$

where

$$P_1(x) = b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}$$

Now if we put $y = x-\xi$ in (3) in place of y, we obtain

$$P(x) = (x-\xi)[A_0(x-\xi)^{n-1} + A_1(x-\xi)^{n-2} + \ldots + A_{n-1}] + P(\xi) \quad (5)$$

Comparing formulas (4) and (5), we conclude that

$$P_1(x) = A_0(x-\xi)^{n-1} + A_1(x-\xi)^{n-2} + \ldots + A_{n-1} \quad (6)$$

whence

$$A_{n-1} = P_1(\xi) \quad (7)$$

Similarly, dividing the polynomial $P_1(x)$ by the binomial $x-\xi$, we can set

$$P_1(x) = (x-\xi)P_2(x) + P_1(\xi) \quad (8)$$

where $P_2(x) = c_0x^{n-2} + c_1x^{n-3} + \ldots + c_{n-2}$.

Besides, from (6) and (7) we have

$$P_1(x) = (x-\xi)[A_0(x-\xi)^{n-2} + A_1(x-\xi)^{n-3} + \ldots + A_{n-2}] + P_1(\xi) \quad (9)$$

Comparing (8) and (9) we conclude that

$$P_2(x) = A_0(x-\xi)^{n-2} + A_1(x-\xi)^{n-3} + \ldots + A_{n-2}$$

whence $A_{n-2} = P_2(\xi)$.

Continuing this process, we successively express all coefficients A_i $(i = 0, 1, \ldots, n)$ in terms of the values of the corresponding polynomials $P_0(x) = P(x)$ and $P_1(x), \ldots, P_n(x) = a_0$ for $x = \xi$:

$$\begin{aligned} A_n &= P(\xi), \\ A_{n-1} &= P_1(\xi), \\ A_{n-2} &= P_2(\xi), \\ &\cdots\cdots \\ A_0 &= P_n(\xi) \end{aligned}$$

where the polynomials $P_{k+1}(x)$ are constructed, proceeding from the polynomials $P_k(x)$, from the formula

$$P_k(x) = (x-\xi)P_{k+1}(x) + P_k(\xi) \qquad (k = 0, 1, \ldots, n)$$

We use the *generalized Horner scheme* for computing the values of $P(\xi)$, $P_1(\xi)$, $P_2(\xi)$, ...:

$$\begin{array}{ccccccc}
a_0 & a_1 & a_2 & \ldots & a_{n-1} & a_n & \underline{|\xi} \\
 & b_0\xi & b_1\xi & \ldots & b_{n-2}\xi & b_{n-1}\xi & \\
\hline
b_0 & b_1 & b_2 & \ldots & b_{n-1} & \underline{b_n = P(\xi)} & \\
 & c_0\xi & c_1\xi & \ldots & c_{n-2}\xi & & \\
\hline
c_0 & c_1 & c_2 & \ldots & \underline{c_{n-1} = P_1(\xi)} & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot &
\end{array}$$

Example. In order to eliminate the term containing the third power of the unknown in the polynomial

$$P(x)=x^4-8x^3+5x^2+2x-7$$

set $x=y+2$.

Find the transformed polynomial.

Solution. Use the generalized Horner scheme:

$$\begin{array}{rrrrr|l}
1 & -8 & 5 & 2 & -7 & \xi=2 \\
 & 2 & -12 & -14 & -24 & \\
\hline
1 & -6 & -7 & -12 & \underline{-31} & \\
 & 2 & -8 & -30 & & \\
\hline
1 & -4 & -15 & \underline{-42} & & \\
 & 2 & -4 & & & \\
\hline
1 & -2 & \underline{-19} & & & \\
 & 2 & & & & \\
\hline
1 & \underline{0} & & & & \\
\hline
\underline{1} & & & & &
\end{array}$$

Hence

$$P(y+2)=y^4-19y^2-42y-31$$

3.3 COMPUTING THE VALUES OF RATIONAL FRACTIONS

Every *rational fraction* $R(x)$ may be represented in the form of a ratio of two polynomials,

$$R(x)=\frac{P(x)}{Q(x)} \tag{1}$$

where

$$P(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n,$$
$$Q(x)=b_0x^m+b_1x^{m-1}+\ldots+b_m$$

Suppose it is required to find the value of $R(x)$ at the point $x=\xi$:

$$R(\xi)=\frac{P(\xi)}{Q(\xi)} \tag{2}$$

The numerator $P(\xi)$ and the denominator $Q(\xi)$ of (2) can be found by means of Horner's scheme. This then yields a simple method for computing the number $R(\xi)$.

Another approach is to transform the function $R(x)$ into a continued fraction. To do this, use the technique described in Sec. 2.3.

3.4 APPROXIMATING THE SUMS OF NUMERICAL SERIES

Definition. The numerical series

$$a_1+a_2+\dots+a_n+\dots \tag{1}$$

is called *convergent* if there exists a limit of the sequence of its partial sums:

$$S=\lim_{n\to\infty} S_n \tag{2}$$

where

$$S_n=a_1+a_2+\dots+a_n$$

The number S is called the *sum of the series.*

Thus, the convergence of series (1) is equivalent to the convergence of the sequence of its partial sums. According to *Cauchy's test* [1] this sequence converges if and only if for an arbitrary $\varepsilon>0$ there is an $N=N(\varepsilon)$ such that

$$|S_{n+p}-S_n|<\varepsilon$$

for $n>N$ and an arbitrary $p>0$.

From formula (2) we get

$$S=S_n+R_n \tag{3}$$

where R_n is the *remainder of the series;* $R_n\to 0$ as $n\to\infty$.

To find the sum S of the convergent series (1) to a specified accuracy ε, choose the number of terms n sufficiently large so that the inequality

$$|R_n|<\varepsilon \tag{4}$$

holds. Then the partial sum S_n can, approximately, be taken for the exact sum S of the series (1).

It will be noted that actually the terms $a_1, a_2, \dots$ are also determined approximately. Besides, the sum S_n is ordinarily rounded off to a given number of decimal places. To take all these errors into account and ensure the required accuracy, one can use the following procedure: in the general case, choose three positive numbers ε_1, ε_2 and ε_3 such that

$$\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$$

Take the number n of terms of the series so large that the *residual error* $|R_n|$ satisfies the inequality

$$|R_n|\leqslant\varepsilon_1 \tag{5}$$

Compute each of the terms a_k $(k=1, 2, \dots, n)$ with a limiting absolute error not exceeding $\frac{\varepsilon_2}{n}$, and let $\bar{a}_k$ $(k=1, 2, \dots, n)$ be the corresponding approximate values of the terms of series (1), that is

$$|\bar{a}_k - a_k| \leqslant \frac{\varepsilon_2}{n}$$

Then for the sum

$$\bar{S}_n = \sum_{k=1}^{n} \bar{a}_k$$

the *error of operation* (summation) satisfies the inequality

$$|S_n - \bar{S}_n| \leqslant \varepsilon_2 \tag{6}$$

Finally, round off the approximate result $\bar{S}_n$ to the simpler number $\bar{\bar{S}}_n$ so that the *rounding error* is

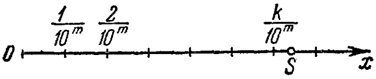
$$|\bar{S}_n - \bar{\bar{S}}_n| \leqslant \varepsilon_3 \tag{7}$$

Then the number $\bar{\bar{S}}_n$ is an approximate value of the sum S of series (1) to the specified accuracy ε. Indeed, from inequalities (5) to (7) we have

$$|S - \bar{\bar{S}}_n| \leqslant |S - S_n| + |S_n - \bar{S}_n| + |\bar{S}_n - \bar{\bar{S}}_n| \leqslant \varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon$$

The number ε is partitioned into the positive parts ε_1, ε_2 and ε_3 in accordance with the volume of work required to obtain the desired result. If $\varepsilon = 10^{-m}$ and the result is to be correct to m decimal places, then one usually takes

Fig. 3



$$\varepsilon_1 = \frac{\varepsilon}{4}, \quad \varepsilon_2 = \frac{\varepsilon}{4}, \quad \varepsilon_3 = \frac{\varepsilon}{2}$$

If there is no final rounding, then one ordinarily sets

$$\varepsilon_1 = \frac{\varepsilon}{2}, \quad \varepsilon_2 = \frac{\varepsilon}{2}, \quad \varepsilon_3 = 0$$

The problem becomes more complicated if one has to find the sum of a series correct, in the narrow meaning of the word, to m decimal places. Geometrically, this means that it is required to find the element of the set $\frac{k}{10^m}$ (k an integer) closest to S (Fig. 3).

Suppose for definiteness the sum S is positive and

$$\tilde{S} = p_0 + \frac{p_1}{10} + \ldots + \frac{p_m}{10^m} + \ldots + \frac{p_n}{10^n}$$

(where p_k are nonnegative integers, $n \geqslant m$) is a rational approximation such that

$$|S - \tilde{S}| \leqslant \frac{1}{10^{m+1}}$$

Suppose that

$$p_{m+1} \neq 4 \text{ and } p_{m+1} \neq 5$$

Then, rounding off the number $\tilde{S}$, we get the desired result:

$$\sigma = p_0 + \frac{p_1}{10} + \ldots + \frac{p_m}{10^m} \quad \text{if } p_{m+1} \leqslant 3 \tag{8}$$

or

$$\sigma = p_0 + \frac{p_1}{10} + \ldots + \frac{p_{m+1}}{10^m} \quad \text{if } p_{m+1} \geqslant 6 \tag{8'}$$

Indeed, in the first case, when rounding down, we have

$$0 \leqslant \tilde{S} - \sigma = \frac{p_{m+1}}{10^{m+1}} + \frac{p_{m+2}}{10^{m+2}} + \ldots + \frac{p_n}{10^n} \leqslant \frac{3}{10^{m+1}} + \frac{9}{10^{m+2}} + \ldots + \frac{9}{10^n} < \frac{4}{10^{m+1}}$$

In the second case, when rounding up, we get

$$0 \leqslant \sigma - \tilde{S} = \frac{1}{10^m} - \frac{p_{m+1}}{10^{m+1}} - \ldots - \frac{p_n}{10^n} \leqslant \frac{1}{10^m} - \frac{6}{10^{m+1}} = \frac{4}{10^{m+1}}$$

Therefore in both cases we have

$$|\tilde{S} - \sigma| \leqslant \frac{4}{10^{m+1}}$$

and so

$$|S - \sigma| \leqslant |S - \tilde{S}| + |\tilde{S} - \sigma| \leqslant \frac{1}{10^{m+1}} + \frac{4}{10^{m+1}} = \frac{1}{2} \cdot 10^{-m}$$

Thus

$$S = \sigma \pm \frac{1}{2} \cdot 10^{-m}$$

If $p_{m+1} = 4$ or $p_{m+1} = 5$, one should increase the accuracy of the approximate sum $\tilde{S}$ by taking another decimal place.

In the particular case when $p_{m+1} = 4$ and it is known that

$$S < \tilde{S}$$

then σ from (8) is an approximate value of the sum S to within $\frac{1}{2} \cdot 10^{-m}$ (too small).

Similarly, if $p_{m+1}=5$ and

$$S>\tilde{S}$$

then σ from (8′) is an approximate value of the sum S to within $\frac{1}{2}\cdot 10^{-m}$ (too large).

To estimate the remainder of the series (1),

$$R_n=a_{n+1}+a_{n+2}+\dots$$

it is useful to apply the following theorems which we give without proof [1].

Theorem 1. *If the terms of the series* (1) *are corresponding values of a positive monotonic decreasing function* f(x), *that is,*

$$a_n=f(n) \quad (n=1,\ 2,\ \dots) \tag{9}$$

then (Fig. 4)

$$\int_{n+1}^{\infty} f(x)\,dx<R_n<\int_{n}^{\infty} f(x)\,dx$$

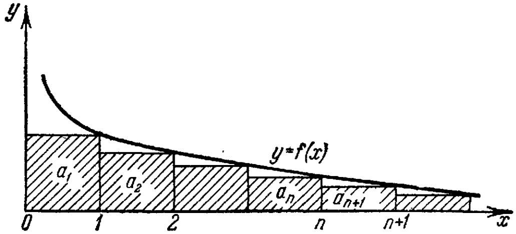


Fig. 4

Theorem 2. *If the series* (1) *is an alternating series:*

$$a_1>0, \quad a_2<0, \quad a_3>0, \ \dots$$

and the moduli of its terms decrease monotonically, then

$$|R_n|\leqslant|a_{n+1}|$$

and

$$\operatorname{sgn} R_n=\operatorname{sgn} a_{n+1}\ ^{1)}$$

Example. Find the sum of the series

$$S=\frac{1}{1^3}+\frac{1}{2^3}+\frac{1}{3^3}+\dots+\frac{1}{n^3}+\dots \tag{10}$$

to within 0.001.

---

1) sgn R_n (the signum function) denotes the *sign* of the number R_n: $\operatorname{sgn} R_n=+1$ if $R_n>0$; $\operatorname{sgn} R_n=-1$ if $R_n<0$; $\operatorname{sgn} R_n=0$ if $R_n=0$.

Solution. We take the residual error as

$$\varepsilon_1 = \frac{1}{4} \cdot 10^{-3} = \frac{1}{4000}$$

The terms of the series (10) are corresponding values of the monotonic decreasing function

$$f(x) = \frac{1}{x^3}$$

And so for the *n*th remainder of the series

$$R_n = \sum_{k=n+1}^{\infty} \frac{1}{k^3}$$

we have the estimate

$$R_n \leqslant \int_n^{\infty} \frac{dx}{x^3} = \frac{1}{2n^2}$$

Solving the inequality

$$\frac{1}{2n^2} \leqslant \frac{1}{4000}$$

we get

$$n \geqslant \sqrt{2000} \approx 44.7$$

We take $n = 45$.

We choose the limiting error of the summation as

$$\varepsilon_2 = \frac{1}{4} \cdot 10^{-3}$$

whence the permissible limiting absolute error of the terms of the partial sum S_{45} of series (10) is

$$\frac{\varepsilon_2}{n} \leqslant \frac{\frac{1}{4} \cdot 10^{-3}}{45} = \frac{5}{9} \cdot 10^{-5}$$

Set

$$\frac{\varepsilon_2}{n} = \frac{1}{2} \cdot 10^{-5}$$

That is to say, we will compute the terms of series (10) correct, in the narrow sense, to five decimal places. Following are the corresponding values of the terms and the results of partial summation:

1.00000	0.00024	0.00003
0.12500	0.00020	0.00003
0.03704	0.00017	0.00003
0.01562	0.00014	0.00003
0.00800	0.00012	0.00002
0.00463	0.00011	0.00002
0.00292	0.00009	0.00002
0.00195	0.00008	0.00002
0.00137	0.00007	0.00002
0.00100	0.00006	0.00002
0.00075	0.00006	0.00001
0.00058	0.00005	0.00001
0.00046	0.00004	0.00001
0.00036	0.00004	0.00001
0.00030	0.00004	0.00001
1.19998	0.00151	0.00029

Thus,

$$S_{45} = 1.19998 + 0.00151 + 0.00029 = 1.20178$$

Rounding this value to thousandths, we get the approximate value of the sum:

$$S \approx 1.202$$

Since the rounding error

$$\varepsilon_3 = 0.00022 < \frac{1}{4} \cdot 10^{-3}$$

the total error of the result does not exceed

$$\frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} < \frac{3}{4} \cdot 10^{-3}$$

Thus,

$$S = 1.202 \pm 0.001$$

A more accurate estimate is obtained if the signs due to rounding are taken into account. By way of comparison we give the value of the sum S to within $\frac{1}{2} \cdot 10^{-6}$ [2]:

$$S = 1.202057$$

Note. Since calculating the total error is an extremely laborious procedure, the practical approach is this: to ensure a given accuracy of $\varepsilon = 10^{-m}$, all intermediate computations are carried out with one or two extra digits. In this process, it is assumed (not quite strictly) that the errors involved do not affect the mth-order decimals of the sought-for result.

It will be noted that in working this example we had to find the sum of a comparatively large number of summands. In practical work, the attempt is made first to transform the series so that the desired result is obtainable with a smaller number of terms. A transformation of this kind is known as *accelerating the convergence of the series* and in many cases affords a great saving in computational time. This problem is considered in Chapter 6.

3.5 COMPUTING THE VALUES OF AN ANALYTIC FUNCTION

A real function $f(x)$ is called analytic at a point ξ if in some neighbourhood $|x-\xi| < R$ of this point the function is expansible in a power series *(Taylor's series)*:

$$f(x) = f(\xi) + f'(\xi)(x-\xi) + \frac{f''(\xi)}{2!}(x-\xi)^2 + \dots$$
$$\dots + \frac{f^{(n)}(\xi)}{n!}(x-\xi)^n + \dots \tag{1}$$

For $\xi = 0$ we get *Maclaurin's series*

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \dots + \frac{f^{(n)}(0)}{n!}x^n + \dots \tag{2}$$

The difference

$$R_n(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(\xi)}{k!}(x-\xi)^k$$

is called the *remainder term* and is the error resulting from the replacement of the function $f(x)$ by the *Taylor polynomial*

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(\xi)}{k!}(x-\xi)^k$$

As is known [1],

$$R_n(x) = \frac{f^{(n+1)}(\xi + \theta(x-\xi))}{(n+1)!}(x-\xi)^{n+1} \tag{3}$$

where $0 < \theta < 1$. In particular, for Maclaurin's series (2) we have [1]

$$R_n(x) = \frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1} \tag{4}$$

where $0 < \theta < 1$. There are also other forms of remainder terms.

In many cases, expanding a function in a Taylor series is a convenient way to compute the values of the function.

If $f(\xi)$ is known and it is required to find the value $f(\xi+h)$, where h is a "small correction", then formula (1) is better written as

$$f(\xi+h)=f(\xi)+f'(\xi)h+\frac{f''(\xi)}{2!}h^2+\ldots+\frac{f^{(n)}(\xi)}{n!}h^n+R_n(h) \quad (5)$$

where

$$R_n(h)=\frac{f^{(n+1)}(\xi+\theta h)}{(n+1)!}h^{n+1} \qquad (0<\theta<1)$$

Example. Approximate $\sqrt{10}$.

Solution. We have

$$\sqrt{10}=\sqrt{3^2+1}=3\left(1+\frac{1}{9}\right)^{\frac{1}{2}} \quad (6)$$

Setting

$$f(x)=(1+x)^{\frac{1}{2}}$$

we successively obtain

$$f'(x)=\frac{1}{2}(1+x)^{-\frac{1}{2}},$$

$$f''(x)=-\frac{1}{4}(1+x)^{-\frac{3}{2}},$$

$$f'''(x)=\frac{3}{8}(1+x)^{-\frac{5}{2}},$$

$$f^{\text{IV}}(x)=-\frac{15}{16}(1+x)^{-\frac{7}{2}}$$

Whence, taking $\xi=0$, $h=\frac{1}{9}$ and noting that

$$f(0)=1, \quad f'(0)=\frac{1}{2}, \quad f''(0)=-\frac{1}{4}, \quad f'''(0)=\frac{3}{8}$$

we find, by virtue of formula (5),

$$\begin{aligned}\left(1+\frac{1}{9}\right)^{\frac{1}{2}}&=1+\frac{1}{2}\cdot\frac{1}{9}-\frac{1}{8}\cdot\frac{1}{81}+\frac{1}{16}\cdot\frac{1}{729}+R_3=\\&=1+0.05556-0.00154+0.00009+R_3=\\&=1.05411+R_3 \qquad (7)\end{aligned}$$

where

$$R_3=-\frac{1}{24}\cdot\frac{15}{16}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}}\cdot\frac{1}{6561}=-\frac{10}{1,680,616}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}}$$

$$(0<\theta<1)$$

Obviously

$$|R_3| < \frac{10}{1{,}680{,}616} < 6 \cdot 10^{-6}$$

From formulas (6) and (7) we get

$$\sqrt{10} = 3.16233 + E \tag{8}$$

where

$$|E| < 3 \cdot \frac{1}{2} \cdot 10^{-5} + 3 \cdot 6 \cdot 10^{-6} = 3.3 \cdot 10^{-5}$$

Rounding off the value obtained to four decimal places, we finally have

$$\sqrt{10} = 3.1623 \pm 6 \cdot 10^{-5}$$

Compare this with the tabular value

$$\sqrt{10} = 3.1622777\ldots$$

3.6 COMPUTING THE VALUES OF EXPONENTIAL FUNCTIONS

For the exponential function e^x we have the expansion [1]

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots \tag{1}$$

whose interval of convergence is $-\infty < x < +\infty$. The remainder term of the series (1) is of the form

$$R_n(x) = \frac{e^{\theta x}}{(n+1)!} x^{n+1} \qquad (0 < \theta < 1) \tag{2}$$

For large absolute values of x, the series (1) is computationally inconvenient. The ordinary procedure is therefore as follows: let

$$x = E(x) + q$$

where $E(x)$ is the largest integer in the number x and $0 \leqslant q < 1$ is the fractional part of the number. We have

$$e^x = e^{E(x)} \cdot e^q \tag{3}$$

The first factor of the product (3) may be found by multiplying:

$$e^{E(x)} = \overbrace{ee \ldots e}^{E(x) \text{ times}} \qquad \text{if } E(x) \geqslant 0$$

or

$$e^{E(x)} = \overbrace{\frac{1}{e} \cdot \frac{1}{e} \cdots \frac{1}{e}}^{-E(x) \text{ times}} \qquad \text{if } E(x) < 0$$

where

$$e = 2.718281828459045\ldots$$

and

$$\frac{1}{e} = 0.367879441171442\ldots$$

Here, in order to ensure the specified accuracy, e or $\frac{1}{e}$ should be taken to a sufficiently large number of decimal places (at the present time the number e has been computed to over 250 decimals).

As for the second factor e^q of the product (3), one can compute it by the expansion given above:

$$e^q = \sum_{n=0}^{\infty} \frac{q^n}{n!} \tag{4}$$

which for $0 \leqslant q < 1$ forms a rapidly converging series, since, on the basis of formula (2), we have for the remainder term $R_n(q)$ the estimate

$$0 \leqslant R_n(q) < \frac{3}{(n+1)!} q^{n+1}$$

Let us now derive a more exact formula for estimating the remainder $R_n(q)$ for $0 < q < 1$. We have

$$\begin{aligned} R_n(q) &= \frac{q^{n+1}}{(n+1)!} + \frac{q^{n+2}}{(n+2)!} + \frac{q^{n+3}}{(n+3)!} + \ldots = \\ &= \frac{q^{n+1}}{(n+1)!} \left[1 + \frac{q}{n+2} + \frac{q^2}{(n+2)(n+3)} + \ldots\right] < \\ &< \frac{q^{n+1}}{(n+1)!} \left[1 + \frac{q}{n+2} + \left(\frac{q}{n+2}\right)^2 + \ldots\right] \end{aligned}$$

From this, by summing the geometric progression in square brackets, we obtain

$$R_n(q) < \frac{q^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{q}{n+2}} \tag{5}$$

or, for $0 < q < 1$, noting that

$$\frac{n+2}{n+1} < \frac{n+1}{n}$$

we finally get

$$0 < R_n(q) < \frac{q^{n+1}}{n!\, n}$$

or

$$0 < R_n(q) < u_n \cdot \frac{q}{n} \tag{6}$$

where $u_n = \frac{q^n}{n!}$ is the last retained term.

If the residual error ε is given, the necessary number of terms n may be found by inspection when solving the inequality

$$\frac{q^{n+1}}{n!\,n} < \varepsilon$$

It is convenient to approximate e^x for small x by formula (1) using the scheme

$$e^x = u_0 + u_1 + u_2 + \ldots + u_n + R_n(x) \tag{7}$$

where

$$u_0 = 1, \quad u_k = \frac{x u_{k-1}}{k} \qquad (k = 1, 2, \ldots, n) \tag{8}$$

On computers, the calculations are conveniently carried out according to the scheme

$$u_k = \frac{x}{k} u_{k-1},$$

$$s_k = s_{k-1} + u_k \qquad (k = 0, 1, 2, \ldots, n)$$

where $u_0 = 1$, $s_{-1} = 0$, $s_0 = 1$. The number $s_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ approximately yields the desired value of e^x.

If ε is the given permissible residual error and $n \geqslant 2|x| > 0$, then the summation process should be terminated as soon as the inequality

$$|R_n(x)| \leqslant R_n(|x|) < \frac{|x|^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{|x|}{n+2}} <$$

$$< \frac{2|x|^{n+1}}{(n+1)!} = \frac{2|x|}{n+1} \cdot \frac{|x|^n}{n!} < |u_n| \leqslant \varepsilon$$

is fulfilled, that is, if

$$|u_n(x)| \leqslant \varepsilon \tag{9}$$

In other words, the summation process is terminated if the last generated term u_n does not exceed ε in modulus, and

$$|R_n(x)| < |u_n|$$

To compute the total error, use the general scheme given in Sec. 3.4.

Example 1. Find $\sqrt{e}$ to within 10^{-5}.

Solution. Assume the residual error to be

$$\varepsilon_1 = \frac{1}{4} \cdot 10^{-5} = 2.5 \cdot 10^{-6}$$

Since, as a rough guess, the number of terms in the sum (7) will then be of the order of ten, we compute the terms to within $\frac{1}{2} \cdot 10^{-7}$, that is to say, to two decimal places.

Setting

$$u_0 = 1, \qquad u_k = \frac{u_{k-1}}{2k} \qquad (k = 1, 2, \ldots)$$

we successively have

$$\left.\begin{aligned}
u_0 &= \quad\ 1 \\
u_1 &= \frac{1}{2} = 0.5000000 \\
u_2 &= \frac{u_1}{4} = 0.1250000 \\
u_3 &= \frac{u_2}{6} = 0.0208333 \\
u_4 &= \frac{u_3}{8} = 0.0026042 \\
u_5 &= \frac{u_4}{10} = 0.0002604 \\
u_6 &= \frac{u_5}{12} = 0.0000217 \\
u_7 &= \frac{u_6}{14} = 0.0000016
\end{aligned}\right\}$$

$$\overline{\quad 1.6487212 \quad}$$

Rounding off the sum to five decimal places, we get

$$\sqrt{e} = 1.64872 \tag{10}$$

with total error

$$\varepsilon < 1.6 \cdot 10^{-6} + 5 \cdot \frac{1}{2} \cdot 10^{-7} + 1.2 \cdot 10^{-6} = 3.05 \cdot 10^{-6} < \frac{1}{2} \cdot 10^{-5}$$

That is, correct (in the narrow sense) to all the digits given in the result (10).

We can also compute e^x by expansion into a continued fraction [4]:

$$e^x = \left[0; \frac{1}{1}, \frac{-2x}{2+x}, \frac{x^2}{6}, \frac{x^2}{10}, \ldots, \frac{x^2}{4n+2}, \ldots\right] \tag{11}$$

which converges for any value of x (real or complex).

Example 2. Find $\sqrt{e}$ by means of formula (11).

Solution. Setting $x=\frac{1}{2}$ in formula (11), construct a table of convergents of the corresponding continued fraction.

k	-1	0	1	2	3	4	5
b_k		0	1	-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
a_k	1	1	1	$\frac{5}{2}$	6	10	14
P_k	1	0	1	$\frac{5}{2}$	$\frac{61}{4}$	$\frac{1225}{8}$	$\frac{34361}{16}$
Q_k	0	1	1	$\frac{3}{2}$	$\frac{37}{4}$	$\frac{743}{8}$	$\frac{20841}{16}$

Stopping with the fifth convergent, we have

$$\sqrt{e}\approx\frac{P_5}{Q_5}=\frac{34361}{16}:\frac{20841}{16}=\frac{34361}{20841}=1.648721$$

to within $\frac{1}{2}\cdot 10^{-6}$.

The following formula may be used to compute the values of the exponential function a^x $(a>0)$:

$$a^x=1+(\ln a)\cdot x+\frac{\ln^2 a}{2!}x^2+\frac{\ln^3 a}{3!}x^3+\dots$$

3.7 COMPUTING THE VALUES OF A LOGARITHMIC FUNCTION

For the natural logarithms of numbers close to unity, the expansion [1]

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\dots$$
$$\dots+(-1)^{n-1}\frac{x^n}{n}+\dots \qquad (-1<x\leqslant 1) \tag{1}$$

holds true. Formula (1) is hardly suitable for computations since the range of numbers $0<1+x\leqslant 2$ is small and, besides, for $|x|$ close to unity, the series (1) converges slowly.

Let us introduce a more convenient formula for computing the natural logarithms of numbers. Replacing x in (1) by $-x$, we have

$$\ln(1-x)=-x-\frac{x^2}{2}-\frac{x^3}{3}-\frac{x^4}{4}-\dots-\frac{x^n}{n}-\dots \tag{2}$$

Subtracting (2) from (1) term by term, we obtain

$$\ln\frac{1-x}{1+x}=-2\left(x+\frac{x^3}{3}+\frac{x^5}{5}+\dots\right)$$

Setting

$$\frac{1-x}{1+x}=z$$

we get

$$x=\frac{1-z}{1+z}$$

and, hence,

$$\ln z=-2\left[\frac{1-z}{1+z}+\frac{1}{3}\left(\frac{1-z}{1+z}\right)^3+\frac{1}{5}\left(\frac{1-z}{1+z}\right)^5+\dots\right] \tag{3}$$

for $0<z<+\infty$.

Let x be a positive number. We represent it as

$$x=2^m\cdot z$$

where m is an integer and $\frac{1}{2}\leqslant z<1$. Then, setting

$$\frac{1-z}{1+z}=\xi$$

where

$$0<\xi\leqslant\frac{1-\frac{1}{2}}{1+\frac{1}{2}}=\frac{1}{3}$$

and using formula (3), we get

$$\ln x=\ln(2^m z)=m\ln 2+\ln z=m\ln 2-2\left(\xi+\frac{\xi^3}{3}+\dots+\frac{\xi^{2n-1}}{2n-1}\right)-R_n$$

where

$$R_n=2\left(\frac{\xi^{2n+1}}{2n+1}+\frac{\xi^{2n+3}}{2n+3}+\frac{\xi^{2n+5}}{2n+5}+\dots\right)<$$
$$<2\cdot\frac{\xi^{2n+1}}{2n+1}(1+\xi^2+\xi^4+\dots)<\frac{2}{1-\xi^2}\cdot\frac{\xi^{2n+1}}{2n+1}$$

For $0<\xi\leqslant\frac{1}{3}$ we have

$$\frac{2}{1-\xi^2}\leqslant\frac{9}{4}$$

and so

$$0<R_n<\frac{9}{4}\cdot\frac{\xi^{2n+1}}{2n+1} \tag{4}$$

or, more crudely,

$$0 < R_n < \frac{1}{4(2n+1)} \cdot \left(\frac{1}{3}\right)^{2n-1}$$

Introducing the notation

$$u_k = \frac{\xi^{2k-1}}{2k-1} \qquad (k = 1,\ 2,\ \ldots)$$

we get

$$\ln x = m \ln 2 - 2(u_1 + u_2 + \ldots + u_n) - R_n \tag{5}$$

where

$$\ln 2 = 0.69314718\ldots$$

The summation process terminates as soon as

$$u_n < 4\varepsilon$$

where ε is the permissible residual error, because then, by formula (4), we have

$$R_n < \frac{9}{4}\xi^2 \cdot \frac{\xi^{2n-1}}{2n-1} \leqslant \frac{1}{4} u_n < \varepsilon$$

The limiting error of the sum $\sum_{k=1}^{n} u_k$ may be estimated by specifying a certain number of decimal places in the summands and establishing, on the basis of (4), the approximate number of summands n.

Example. Find $\ln 3$ to within 10^{-5}.

Solution. We will perform the computations with two additional digits. Putting

$$3 = 2^2 \cdot \frac{3}{4} = 2^2 \cdot 0.75$$

we have $z = 0.75$ and

$$\xi = \frac{1-z}{1+z} = \frac{0.25}{1.75} = \frac{1}{7} = 0.1428571$$

We have

$$\left.\begin{aligned} u_1 &= \xi = 0.1428571 \\ u_2 &= \frac{\xi^3}{3} = 0.0009718 \\ u_3 &= \frac{\xi^5}{5} = 0.0000119 \\ u_4 &= \frac{\xi^7}{7} = 0.0000002 \end{aligned}\right\}$$

$$\overline{\quad 0.1438410 \quad}$$

Using formula (5), we get

$$\ln 3 = 2\cdot 0.69314718 - 2\cdot 0.1438410 = 1.09861$$

Note. It is also possible to compute natural logarithms by proceeding from the representation

$$x = e^{p}z$$

where p is an integer and $\frac{1}{e} < z \leqslant 1$ (see [5]).

To compute common logarithms, use the formula

$$\log_{10} x = M \ln x$$

where

$$M = \log_{10} e = 0.434294481903252\ldots$$

3.8 COMPUTING THE VALUES OF TRIGONOMETRIC FUNCTIONS

A. COMPUTING SINE AND COSINE VALUES

Using reduction formulas it is possible to confine the argument x to the interval $0 \leqslant x \leqslant \frac{\pi}{2}$. If $0 \leqslant x \leqslant \frac{\pi}{4}$, we have

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \tag{1}$$

but if $\frac{\pi}{4} \leqslant x \leqslant \frac{\pi}{2}$, then we set

$$\sin x = \cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} \tag{2}$$

where $z = \frac{\pi}{2} - x$ and $0 \leqslant z \leqslant \frac{\pi}{4}$.

The sum of the series (1) may conveniently be computed by the summation process

$$\sin x = u_1 + u_2 + \ldots + u_n + R_n \tag{3}$$

where the summands u_k $(k = 1, 2, \ldots, n)$ are successively found by means of the recurrence relation

$$u_1 = x,\ u_{k+1} = -\frac{x^2}{2k(2k+1)} u_k \qquad (k = 1, 2, \ldots, n-1)$$

Since the series (1) is an alternating series with terms monotonically decreasing in modulus, for the remainder term R_n the estimate

$$|R_n| \leqslant \frac{x^{2n+1}}{(2n+1)!} = |u_{n+1}|$$

holds true, and

$$\operatorname{sgn} R_n = \operatorname{sgn} u_{n+1}$$

And we can terminate the summation process as soon as we find that

$$|u_n| \leqslant \varepsilon$$

where ε is the specified residual error.

Analogously

$$\cos z = v_1 + v_2 + \ldots + v_n + R_n$$

where

$$v_1 = 1, \quad v_{k+1} = -\frac{x^2}{(2k-1)\,2k} v_k \qquad (k = 1, 2, \ldots, n-1)$$

and

$$|R_n| \leqslant \frac{z^{2n}}{(2n)!} = |v_{n+1}|, \quad \operatorname{sgn} R_n = \operatorname{sgn} v_{n+1}$$

Example. Find $\sin 20°30'$ to within 10^{-5}.

Solution. We have

$$x = \operatorname{arc} 20°30' = \frac{\pi}{9} + \frac{\pi}{360} = 0.349066 + 0.008727 = 0.357793$$

Using formula (3), we obtain

$$\left.\begin{aligned} u_1 &= \ \ x \ \ = 0.357793 \\ u_2 &= \frac{x^2 u_1}{2\cdot 3} = -0.007634 \\ u_3 &= \frac{x^2 u_2}{4\cdot 5} = +0.000049 \\ u_4 &= \frac{x^2 u_3}{6\cdot 7} = -0.000000 \end{aligned}\right\}$$

$$\overline{\quad -0.350208 \quad}$$

whence

$$\sin 20°30' = 0.35021$$

The values of $\cos x$ are computed in a similar manner.

B. COMPUTING TANGENTS

We can take it that $0 \leqslant x \leqslant \frac{\pi}{4}$. For $\tan x$, when $|x| < \frac{\pi}{2}$, the following expansion [6] holds true:

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots$$

The coefficients of the expansion are expressed in terms of Bernoulli numbers (see Sec. 16.12).

The value of a tangent is very simply computed by means of continued fractions. Assuming

$$\tan x = \frac{x}{y}$$

we have, by virtue of Lambert's formula (Sec. 2.6),

$$y=\left[1;\ \frac{-x^2}{3},\ \frac{-x^2}{5},\ \ldots,\ \frac{-x^2}{2n+1},\ \ldots\right]$$

In order to compute y accurate to 10^{-10} it is sufficient to take $n=7$. Then

$$y=1-x^2:(3-x^2:(5-x^2:(7-x^2:(9-x^2:(11-\\ -x^2:(13-x^2:15)))))) \qquad (4)$$

Ordinarily, y is computed with the aid of the Horner scheme for division (beginning at the end):

$$\begin{aligned} y_1 &= 13-x^2:15,\\ y_2 &= 11-x^2:y_1,\\ y_3 &= 9-x^2:y_2,\\ y_4 &= 7-x^2:y_3,\\ y_5 &= 5-x^2:y_4,\\ y_6 &= 3-x^2:y_5,\\ y=y_7 &= 1-x^2:y_6 \end{aligned}$$

Whence $\tan x=\frac{x}{y}$.

Example. Find tan 40°.

Solution. We have

$$x=\text{arc}\,40^\circ=0.698132$$

and

$$x^2=0.487388$$

From this,

$$\begin{aligned} y_1 &= 13-\frac{0.487388}{15}=12.967508,\\ y_2 &= 11-\frac{0.487388}{12.967508}=10.962413,\\ y_3 &= 9-\frac{0.487388}{10.962413}=8.955540,\\ y_4 &= 7-\frac{0.487388}{8.955540}=6.955577,\\ y_5 &= 5-\frac{0.487388}{6.955577}=4.929928,\\ y_6 &= 3-\frac{0.487388}{4\ 929928}=2.901137,\\ y=y_7 &= 1-\frac{0.487388}{2.901137}=0.832001 \end{aligned}$$

and so

$$\tan 40^\circ=\frac{0.698132}{0.832001}=0.839100$$

All digits are correct in this result.

3.9 COMPUTING THE VALUES OF HYPERBOLIC FUNCTIONS

A. COMPUTING THE VALUES OF THE HYPERBOLIC SINE

We know that

$$\sinh x = \frac{e^x - e^{-x}}{2}$$

and

$$\sinh(-x) = -\sinh x$$

The following expansion holds for the hyperbolic sine:

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots \qquad (-\infty < x < +\infty)$$

Assuming $x > 0$, the computations are conveniently performed by the summation process:

$$\sinh x = u_1 + u_2 + \dots + u_n + R_n$$

where

$$u_1 = x, \quad u_{k+1} = \frac{x^2}{2k(2k+1)} u_k \qquad (k = 1, 2, \dots, n-1)$$

and R_n is the remainder term. For $n \geqslant x > 0$ we have

$$R_n = \frac{x^{2n+1}}{(2n+1)!} + \frac{x^{2n+3}}{(2n+3)!} + \frac{x^{2n+5}}{(2n+5)!} + \dots <$$

$$< \frac{x^{2n+1}}{(2n+1)!}\left[1 + \frac{x^2}{(2n+2)(2n+3)} + \frac{x^4}{(2n+2)^2(2n+3)^2} + \dots\right] <$$

$$< \frac{x^{2n+1}}{(2n+1)!} \cdot \frac{1}{1 - \frac{x^2}{(2n+2)(2n+3)}} < \frac{4}{3}\frac{x^{2n+1}}{(2n+1)!} = \frac{4}{3} u_{n+1}$$

Since, obviously,

$$u_{n+1} = \frac{x^2}{2n(2n+1)} u_n < \frac{1}{4} u_n$$

it follows that

$$R_n < \frac{1}{3} u_n$$

B. COMPUTING THE VALUES OF THE HYPERBOLIC COSINE

As we know,

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

and

$$\cosh(-x) = \cosh x$$

The following expansion holds true for the hyperbolic cosine:

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots \quad (-\infty < x < +\infty)$$

The computations are conveniently carried out by the summation process

$$\cosh x = v_1 + v_2 + \ldots + v_n + R_n$$

where

$$v_1 = 1, \; v_{k+1} = \frac{x^2}{(2k-1)\,2k} v_k \qquad (k = 1, 2, \ldots, n-1)$$

and R_n is the remainder term. For $n \geqslant |x| > 0$ we have

$$R_n = \frac{x^{2n}}{(2n)!} + \frac{x^{2n+2}}{(2n+2)!} + \frac{x^{2n+4}}{(2n+4)!} + \ldots <$$
$$< \frac{x^{2n}}{(2n)!}\left[1 + \frac{x^2}{(2n+1)(2n+2)} + \frac{x^4}{(2n+1)^2(2n+2)^2} + \ldots\right] <$$
$$< \frac{x^{2n}}{(2n)!} \cdot \frac{1}{1 - \dfrac{x^2}{(2n+1)(2n+2)}} < \frac{4}{3} \cdot \frac{x^{2n}}{(2n)!} = \frac{4}{3} v_{n+1}$$

Since for $n \geqslant 1$ the inequality

$$v_{n+1} = \frac{x^2}{(2n-1)\,2n} v_n \leqslant \frac{1}{2} v_n$$

holds true, it follows that

$$R_n < \frac{2}{3} v_n$$

C. COMPUTING THE HYPERBOLIC TANGENT

We have

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

where

$$\tanh(-x) = -\tanh x$$

For $|x|$ small, the following expansion can be used to compute the value of the hyperbolic tangent:

$$\tanh x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots \quad \left(|x| < \frac{\pi}{2}\right).$$

For any x, the value of the hyperbolic tangent is expressed by a continued fraction:

$$\tanh x = \left[0; \frac{x}{1}, \frac{x^2}{3}, \frac{x^2}{5}, \ldots, \frac{x^2}{2n-1}, \ldots\right]$$

And since for $x>0$ the elements of this fraction are positive, $\tanh x$, for $x>0$, lies between two successive convergents.

If $x>0$ is great, then $\tanh x$ is conveniently computed by the formula

$$\tanh x = 1 - \frac{2}{e^{2x}+1}$$

3.10. USING THE METHOD OF ITERATION FOR APPROXIMATING THE VALUES OF A FUNCTION

Let it be required to compute the value of the continuous function

$$y=f(x) \tag{1}$$

for a given value of the argument x. If the function (1) is complicated and one has to compute a large number of its values, the computations are ordinarily performed on computers. It may happen that direct computation of the values of the function by formula (1) will be difficult due to the design features of the machine. The simplest operations may turn out to be "complicated" and even impossible to perform. For instance, there are calculating machines that do not perform division. In such cases the following technique is often useful. Write (1) in implicit form:

$$F(x, y)=0 \tag{2}$$

We assume that $F(x, y)$ is continuous and has a continuous partial derivative $F'_y(x, y) \neq 0$.

Let y_n be an approximate value of y. Using Lagrange's theorem, we have

$$F(x, y_n) = F(x, y_n) - F(x, y) = (y_n - y) F'_y(x, \bar{y}_n)$$

where $\bar{y}_n$ is an intermediate value between y_n and y, whence

$$y = y_n - \frac{F(x, y_n)}{F'_y(x, \bar{y}_n)} \tag{3}$$

We do not know the value of $\bar{y}_n$. Assuming $\bar{y}_n \approx y_n$, we obtain the following *iterative process* [7] for computing the value of y:

$$y_{n+1} = y_n - \frac{F(x, y_n)}{F'_y(x, y_n)} \qquad (n=0, 1, 2, \ldots) \tag{4}$$

Formula (3) has a simple geometric meaning. Fix the value of x and consider the graph of the function

$$z=F(x, y) \tag{4'}$$

From formula (4) it follows that our process is the *Newton method* (see Sec. 4.5) applied to (4); that is, the successive approximations y_{n+1} are obtained as the abscissas of the points of intersection with the y-axis of the tangent to the curve (4) drawn at $y=y_n$ $(n=0, 1, 2, \ldots)$ (Fig. 5). The convergence of the process is ensured if

$$F'_y(x, y) \quad \text{and} \quad F''_{yy}(x, y)$$

retain constant signs in the interval under consideration that contains the root of y.

Generally speaking, the initial value y_0 is arbitrary and is chosen as close as possible to the desired value of y. The process of iteration is continued until, within the limits of the given accuracy ε, two successive values y_n and y_{n-1} coincide: $|y_{n-1}-y_n| < \varepsilon$. Strictly speaking, there is no guarantee here that

$$|y-y_n| < \varepsilon \tag{5}$$

For this reason, each concrete case requires an additional investigation.

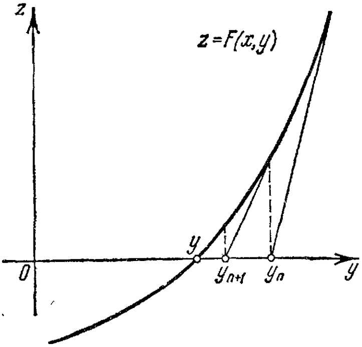


Fig. 5

The merit of iterative processes is that the operations are of the same type and therefore are comparatively easy to programme.

It is worth noting that the representation $F(x, y)=0$ for the given function (1) may be realized in an infinitude of ways. This fact should be utilized in order to obtain a rapidly converging iteration process. Types of the basic processes are given in the sections which follow.

3.11 COMPUTING RECIPROCALS

Let $y=\frac{1}{x}$.

For definiteness we assume that $x>0$. Set

$$F(x, y) \equiv x-\frac{1}{y}=0$$

Then

$$F'_y(x, y)=\frac{1}{y^2}$$

Using formula (4) of Sec. 3.10, we have

$$y_{n+1}=y_n-\frac{x-\frac{1}{y_n}}{\frac{1}{y_n^2}}$$

or

$$y_{n+1} = y_n (2 - x y_n) \qquad (n = 0, 1, 2, \ldots) \tag{1}$$

We have thus obtained an iterative process without division. The initial value y_0 is chosen in the following manner. Suppose the argument is written in binary:

$$x = 2^m x_1, \text{ where } m \text{ is an integer and } \frac{1}{2} \leqslant x_1 < 1$$

Then set

$$y_0 = 2^{-m} \tag{2}$$

Let us examine the conditions of convergence of process (1). From formula (1) we have

$$\frac{1}{x} - y_n = \frac{1}{x} - 2y_{n-1} + x y^2_{n-1} = x \left(\frac{1}{x} - y_{n-1} \right)^2 \tag{3}$$

whence

$$\frac{1}{x} - y_n = x^{2^n - 1} \left(\frac{1}{x} - y_0 \right)^{2^n} = \frac{1}{x} (1 - x y_0)^{2^n} \tag{4}$$

For the convergence of process (4), it is necessary and sufficient that the inequality

$$|1 - x y_0| < 1$$

hold or that

$$-1 < 1 - x y_0 < 1$$

Thus, we finally get the following result: if

$$0 < x y_0 < 2 \tag{5}$$

then

$$\lim_{n \to \infty} y_n = \frac{1}{x}$$

Note that for our choice of y_0 in (2), we have

$$x y_0 = 2^m x_1 \cdot 2^{-m} = x_1$$

and so

$$\frac{1}{2} \leqslant x y_0 < 1 \tag{6}$$

and hence condition (5) is met. Besides, we conclude from formula (3) that

$$\left| \frac{1}{x} - y_n \right| \leqslant \frac{1}{x} \left(\frac{1}{2} \right)^{2^n} \leqslant 2 y_0 \left(\frac{1}{2} \right)^{2^n}$$

That is to say, the convergence of the iteration process is extremely rapid.

Let us derive a different estimate of the error in the value of y_n which is sometimes of more practical convenience. Note first of all that the successive approximations y_0, y_1, y_2, ... are obtained in the given case by Newton's method as applied to the hyperbola

$$z = x - \frac{1}{y} \qquad (x = \text{const})$$

(Fig. 6). From the inequality (6) and formula (3) it follows that

$$0 < y_n < \frac{1}{x} \qquad (n = 0,\ 1,\ 2,\ \ldots)$$

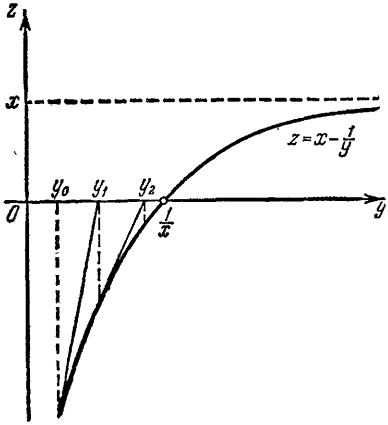


Fig. 6

Besides, since

$$y_n - y_{n-1} = y_{n-1}(1 - xy_{n-1}) = \\ = xy_{n-1}\left(\frac{1}{x} - y_{n-1}\right) \geqslant 0 \qquad (7)$$

it follows that the successive approximations of y_n increase monotonically:

$$y_0 \leqslant y_1 \leqslant y_2 \leqslant \ldots$$

From formula (7) we have

$$\frac{1}{x} - y_{n-1} = \frac{1}{xy_{n-1}}(y_n - y_{n-1})$$

or, since

$$xy_{n-1} \geqslant xy_0 \geqslant \frac{1}{2}$$

it follows that

$$\frac{1}{x} - y_{n-1} \leqslant 2(y_n - y_{n-1})$$

Whence

$$\frac{1}{x} - y_n \leqslant y_n - y_{n-1}$$

Thus, if it is found that $y_n - y_{n-1} < \varepsilon$, then the true error too is

$$0 < \frac{1}{x} - y_n < \varepsilon$$

Example. Using (1) find the value of the function $y = \frac{1}{x}$ for $x = 3$.

Solution. Here, $x = 2^2 \cdot \frac{3}{4}$. Putting $y_0 = \frac{1}{4}$, we have

$$y_1 = \frac{1}{4}\left(2 - \frac{3}{4}\right) = \frac{5}{16} = 0.312,$$
$$y_2 = 0.312\,(2 - 3 \cdot 0.312) = 0.332, \text{ etc.}$$

The iteration process converges rapidly.

3.12 COMPUTING SQUARE ROOTS

Let

$$y=\sqrt{x} \qquad (x>0) \tag{1}$$

Set

$$F(x,\ y)\equiv y^2-x=0$$

and then

$$F'_y(x,\ y)=2y$$

Using formula (4) of Sec. 3.10, we have

$$y_{n+1}=y_n-\frac{y_n^2-x}{2y_n}$$

or

$$y_{n+1}=\frac{1}{2}\left(y_n+\frac{x}{y_n}\right) \tag{2}$$

($n=0,\ 1,\ 2,\ \ldots$) which is *Hero's process* (*Hero's algorithm*).

The successive approximations $y_0,\ y_1,\ y_2,\ \ldots$ are clearly obtained by Newton's method applied to the parabola

$$z=y^2-x \quad (x=\text{const})$$

(Fig. 7).

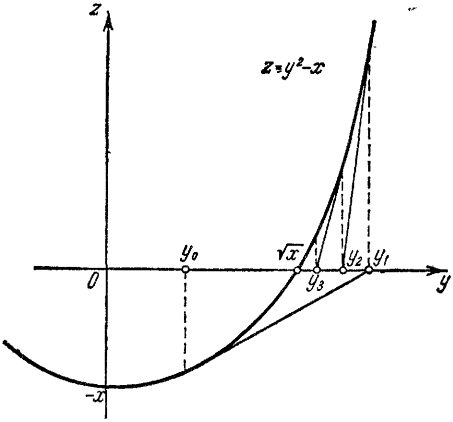


Fig. 7

Note that if for y_0 we take the tabular value, which gives $\sqrt{x}$ with a relative error $|\delta|$, then y_1 determined from (2) will yield the value of $\sqrt{x}$ with, approximately, the relative error $\frac{1}{2}\delta^2$.

Indeed, setting

$$y_0=\sqrt{x}\,(1+\delta)$$

and neglecting powers of δ above the third, we have

$$y_1=\frac{1}{2}\left(y_0+\frac{x}{y_0}\right)=\frac{1}{2}\left[\sqrt{x}\,(1+\delta)+\sqrt{x}\,(1+\delta)^{-1}\right]=$$
$$=\frac{1}{2}\sqrt{x}\,(1+\delta+1-\delta+\delta^2)=\sqrt{x}\left(1+\frac{\delta^2}{2}\right)$$

From this we draw the following important conclusion: *when Hero's process is used, the number of correct digits is roughly doubled at each step compared to the original number of correct digits.*

Example 1. For $y=\sqrt{2}$ we have approximately

$$y_0 = 1.4$$

Making this value more precise, we obtain

$$y_1=\frac{1}{2}\left(1.4+\frac{2}{1.4}\right)=0.7+0.714=1.414$$

Repeating the process, we get

$$y_2=\frac{1}{2}\left(1.414+\frac{2}{1.414}\right)=0.707+0.7072136=1.4142136$$

correct to eight or seven decimal places. Indeed,

$$\sqrt{2}=1.41421356\ldots$$

Let us examine the conditions of convergence of the Hero process. From formula (2), replacing $n+1$ by n, we have, for $y_0 \neq 0$,

$$y_n-\sqrt{x}=\frac{1}{2y_{n-1}}(y_{n-1}-\sqrt{x})^2$$

and

$$y_n+\sqrt{x}=\frac{1}{2y_{n-1}}(y_{n-1}+\sqrt{x})^2$$

whence

$$\frac{y_n-\sqrt{x}}{y_n+\sqrt{x}}=\left(\frac{y_{n-1}-\sqrt{x}}{y_{n-1}+\sqrt{x}}\right)^2 \tag{3}$$

Consequently

$$\frac{y_n-\sqrt{x}}{y_n+\sqrt{x}}=\left(\frac{y_0-\sqrt{x}}{y_0+\sqrt{x}}\right)^{2^n}$$

and

$$y_n-\sqrt{x}=2\sqrt{x}\cdot\frac{q^{2^n}}{1-q^{2^n}} \tag{4}$$

where

$$q=\frac{y_0-\sqrt{x}}{y_0+\sqrt{x}} \tag{5}$$

From formula (4) it follows that the Hero process converges for

$$|q|<1$$

that is, if

$$y_0>0$$

In this case we clearly have

$$\lim_{n\to\infty} y_n=\sqrt{x}$$

also

$$y_n \geqslant \sqrt{x} \qquad (n = 1, 2, \ldots)$$

Note that

$$y_{n-1} - y_n = y_{n-1} - \frac{1}{2}\left(y_{n-1} + \frac{x}{y_{n-1}}\right) = \frac{y_{n-1}^2 - x}{2y_{n-1}} > 0 \tag{6}$$

and so the approximations y_n for $n \geqslant 1$ form a monotonic decreasing sequence

$$y_1 \geqslant y_2 \geqslant \ldots \geqslant y_{n-1} \geqslant y_n \geqslant \ldots \geqslant \sqrt{x}$$

(Equality can only occur when $y_0 = \sqrt{x}$.)

When doing the computation on a computer, it is convenient to write the number x in binary: $x = 2^m x_1$, where m is an integer and $\frac{1}{2} \leqslant x_1 < 1$. Then for the zero approximation one ordinarily takes

$$y_0 = 2^{E\left(\frac{m}{2}\right)} \tag{7}$$

where $E\left(\frac{m}{2}\right)$ denotes the greatest integer in the number $\frac{m}{2}$.

Example 2. Find $\sqrt{5}$.

Solution. Here $x = 5 = 2^3 \cdot \frac{5}{8}$, and so

$$y_0 = 2^{E\left(\frac{3}{2}\right)} = 2$$

Using formula (2) we successively find

$$y_1 = \frac{1}{2}\left(2 + \frac{5}{2}\right) = 2.25,$$

$$y_2 = \frac{1}{2}\left(2.25 + \frac{5}{2.25}\right) = \frac{1}{2}(2.25 + 2.2222) = 2.2361$$

and so on. Using square-root tables we have

$$\sqrt{5} = 2.236068\ldots$$

Let us estimate the quantity $|q|$ expressed by formula (5) by proceeding from the value of y_0 as determined by (7).

If $m = 2p$ is even, then

$$y_0 = 2^{E\left(\frac{m}{2}\right)} = 2^p > \sqrt{x}$$

and, hence,

$$|q| = \frac{y_0 - \sqrt{x}}{y_0 + \sqrt{x}} = \frac{2^p - 2^p\sqrt{x_1}}{2^p + 2^p\sqrt{x_1}} = \frac{1 - \sqrt{x_1}}{1 + \sqrt{x_1}} \leqslant \frac{1 - \sqrt{\frac{1}{2}}}{1 + \sqrt{\frac{1}{2}}} = (\sqrt{2} - 1)^2$$

Analogously, if $m=2p+1$ is odd, then

$$y_0=2^{E\left(\frac{m}{2}\right)}=2^p\leqslant\sqrt{x}$$

Therefore

$$|q|=\frac{\sqrt{x}-y_0}{\sqrt{x}+y_0}=\frac{2^p\sqrt{2x_1}-2^p}{2^p\sqrt{2x_1}+2^p}=\frac{\sqrt{2x_1}-1}{\sqrt{2x_1}+1}=$$
$$=1-\frac{2}{\sqrt{2x_1}+1}<1-\frac{2}{\sqrt{2}+1}=(\sqrt{2}-1)^2$$

We thus always have

$$|q|\leqslant(\sqrt{2}-1)^2=0.1716\ldots<\frac{1}{5}$$

Whence, on the basis of (4), we get

$$0\leqslant y_n-\sqrt{x}<2\sqrt{x}\frac{\left(\frac{1}{5}\right)^{2^n}}{1-\left(\frac{1}{5}\right)^{2^n}}\leqslant\frac{25}{12}y_1\left(\frac{1}{5}\right)^{2^n}\quad\text{for}\quad n\geqslant 1$$

where

$$y_1=\frac{1}{2}\left(y_0+\frac{x}{y_0}\right)\leqslant\frac{3}{2}y_0$$

And from this,

$$0\leqslant y_n-\sqrt{x}<\frac{25}{8}y_0\left(\frac{1}{5}\right)^{2^n}\tag{8}$$

From (8), it is easy to determine the number of iterations $n=n(x)$ sufficient to ensure a given accuracy.

We give one more formula for estimating the error in the value of $y_n\ (n\geqslant 2)$. Since

$$y_{n-1}\geqslant\sqrt{x}\quad\text{and}\quad\frac{x}{y_{n-1}}\leqslant\sqrt{x}$$

we have, taking into account (6),

$$y_{n-1}-\sqrt{x}\leqslant y_{n-1}-\frac{x}{y_{n-1}}=\frac{y_{n-1}^2-x}{y_{n-1}}=2(y_{n-1}-y_n)$$

Hence

$$0\leqslant y_n-\sqrt{x}\leqslant y_{n-1}-y_n\tag{9}$$

Thus, if $0\leqslant y_{n-1}-y_n<\varepsilon\ (n\geqslant 2)$ it is guaranteed that $0\leqslant y_n-\sqrt{x}<\varepsilon$.

Another method for computing square roots that is sometimes useful is this. Replace the function (1) by the equivalent relation

$$F(x,\ y)\equiv\frac{x}{y^2}-1=0$$

Then

$$F'_y(x, y) = -\frac{2x}{y^3}$$

Using formula (4) of Sec. 3.10, we get

$$y_{n+1} = y_n + \frac{\frac{x}{y_n^2} - 1}{\frac{2x}{y_n^3}}$$

or

$$y_{n+1} = \frac{y_n}{2}\left(3 - \frac{y_n^2}{x}\right) \qquad (n = 0, 1, 2, \ldots) \tag{10}$$

We will not consider the error estimate or the conditions for convergence of the iterative process (10).

3.13 COMPUTING THE RECIPROCAL OF A SQUARE ROOT

Set

$$y = \frac{1}{\sqrt{x}} \quad (x > 0)$$

Writing the function as

$$y = \sqrt{\frac{1}{x}}$$

we obtain from (10) of the preceding section an iterative process without division:

$$y_{n+1} = \frac{y_n}{2}(3 - xy_n^2) \qquad (n = 0, 1, 2, \ldots) \tag{1}$$

If $x = 2^m x_1$, where $\frac{1}{2} \leqslant x_1 < 1$, then for y_0 choose the value

$$y_0 = 2^{-E\left(\frac{m}{2}\right)}$$

It will be noted that by using the obvious equation

$$\sqrt{x} = x\sqrt{\frac{1}{x}}$$

it is possible, by virtue of formula (1), to extract the square root of a number without invoking the operation of division.

3.14 COMPUTING CUBE ROOTS

If

$$y = \sqrt[3]{x} \qquad (x > 0) \tag{1}$$

then, putting

$$F(x, y) \equiv y^3 - x = 0$$

we get

$$F'_y(x, y) = 3y^2$$

whence, using formula (4) of Sec. 3.10, we obtain

$$y_{n+1} = y_n - \frac{y_n^3 - x}{3y_n^2} \tag{2}$$

or

$$y_{n+1} = \frac{1}{3}\left(2y_n + \frac{x}{y_n^2}\right) \tag{3}$$

Geometrically, process (3) is Newton's method applied to the cubic parabola

$$z = y^3 - x \quad (x = \text{const})$$

(Fig. 8). The process (3) converges for $y_0 > 0$.

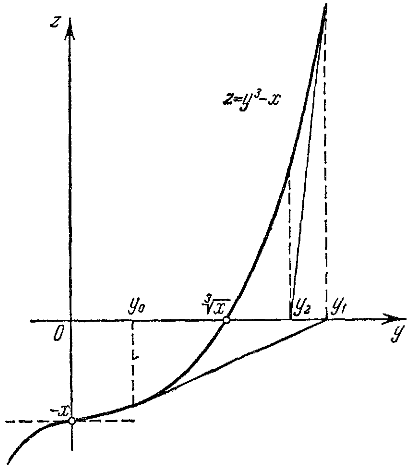


Fig. 8

If for the initial approximation y_0 we take the tabular value of $\sqrt[3]{x}$ which has a relative error of $|\delta|$, that is, if we set

$$y_0 = \sqrt[3]{x}(1 + \delta)$$

then the value of y_1 found from formula (3) will yield $\sqrt[3]{x}$ with a relative error of δ^2. Indeed, applying formula (3), we have

$$y_1 = \frac{1}{3}\left(2y_0 + \frac{x}{y_0^2}\right) = \frac{1}{3}\left[2\sqrt[3]{x}(1+\delta) + \sqrt[3]{x}(1+\delta)^{-2}\right] =$$
$$= \frac{1}{3}\sqrt[3]{x}(2 + 2\delta + 1 - 2\delta + 3\delta^2) = \sqrt[3]{x}(1 + \delta^2)$$

From this we conclude, for one thing, that if y_0 is correct to p digits, in the narrow sense, then y_1 will be correct to $2p$ or $2p-1$ digits, in the broad sense (cf. Sec. 3.12).

Example. Using three-place tables, we have

$$\sqrt[3]{10} = 2.154$$

correct to all the digits.

Using formula (3), we get

$$\sqrt[3]{10}=\frac{1}{3}\left(2\cdot 2.154+\frac{10}{2.154^2}\right)=\frac{1}{3}(2\cdot 2.154+2.155304)=2.154435$$

By way of comparison, Barlow's tables give

$$\sqrt[3]{10}=2.1544347\ldots$$

If $x=2^m x_1$, where m is an integer and $\frac{1}{2}\leqslant x_1<1$, then

$$y_0=2^{E\left(\frac{m}{3}\right)}>0 \tag{4}$$

is ordinarily taken for the initial value y_0.

Since

$$y_n-\sqrt[3]{x}=\frac{1}{3}\left(2y_{n-1}+\frac{x}{y_{n-1}^2}-3\sqrt[3]{x}\right)=$$
$$=\frac{1}{3y_{n-1}^2}(y_{n-1}-\sqrt[3]{x})^2(2y_{n-1}+\sqrt[3]{x})>0$$

it follows that

$$y_n\geqslant\sqrt[3]{x} \quad \text{for} \quad n\geqslant 1 \tag{5}$$

Besides, from formula (2), replacing $n+1$ by n, we have

$$y_{n-1}-y_n=\frac{y_{n-1}^3-x}{3y_{n-1}^2} \tag{6}$$

and therefore

$$y_1\geqslant y_2\geqslant\ldots\geqslant y_{n-1}\geqslant y_n\geqslant\ldots\geqslant\sqrt[3]{x} \tag{7}$$

whence it follows that the limit

$$\lim_{n\to\infty} y_n=y>0$$

exists. Passing to the limit in (3) as $n\longrightarrow\infty$, we get

$$y=\frac{1}{3}\left(2y+\frac{x}{y^2}\right)$$

That is, $y^3=x$, hence, $y=\sqrt[3]{x}$. Thus

$$\lim_{n\to\infty} y_n=\sqrt[3]{x}$$

If the initial approximation y_0 is chosen on the basis of formula (4), it can be proved that

$$0\leqslant y_n-\sqrt[3]{x}\leqslant\frac{3}{2}(y_{n-1}-y_n)$$

for $n\geqslant 2$.

REFERENCES FOR CHAPTER 3

[1] *V. I. Smirnov, Course of Higher Mathematics*, 1957, Vol. I, Chapter IV (in Russian).

[2] *A. Markov, Calculus of Finite Differences*, 1911, Chapter III (in Russian).

[3] *G. P. Tolstov, Course of Mathematical Analysis*, 1957, Vol. II, Chapter XXIV (in Russian)

[4] *A. N. Khovansky, Application of Continued Fractions and Their Generalizations to Problems of Approximate Analysis*, 1956, Chapter II (in Russian).

[5] *B. M. Kagan and T. M. Ter-Mikaelyan, Solution of Engineering Problems on Automatic Digital Computers*, 1958, Chapter III (in Russian).

[6] *G. M. Fikhtengolts, Course of Differential and Integral Calculus*, 1948, Vol. II, Chapter XII (in Russian).

[7] *L. A. Lyusternik, A. A. Abramov, V. I. Shestakov, M. R. Shura-Bura, The Solution of Mathematical Problems on Automatic Digital Computers*, 1952 (in Russian).

Chapter 4

APPROXIMATE SOLUTIONS OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

4.1 ISOLATION OF ROOTS

If an algebraic or transcendental equation is fairly complicated, it is not usually possible to find exact roots. What is more, in certain cases the equation contains coefficients known only approximately and, hence, the very aim of finding the exact roots of the equation is meaningless. Therefore, methods for approximating the roots of an equation and estimating their degree of accuracy acquire particular importance.

Suppose we have an equation

$$f(x) = 0 \tag{1}$$

where the function $f(x)$ is defined and continuous on some finite or infinite interval $a < x < b$.

In certain cases in the sequel we will need the existence and continuity of the first derivative $f'(x)$ or even the second derivative $f''(x)$. This will be appropriately stipulated where necessary.

Every value ξ for which the function $f(x)$ is zero, that is, such that

$$f(\xi) = 0$$

is called a *root of equation* (1) or a *zero* of the function $f(x)$.

We will assume that equation (1) has only *isolated roots*, that is, for each root of (1) there is a neighbourhood which does not contain any other roots of the equation.

Approximating the isolated real roots of (1) ordinarily consists of two stages:

(1) **isolating the roots,** that is, establishing the smallest possible intervals $[\alpha, \beta]$ containing one and only one root of equation (1);

(2) **improving the values of the approximate roots,** that is, refining them to the specified degree of accuracy.

Very useful in the isolating of roots is the following familiar theorem of mathematical analysis ([5], Chapter 4).

Theorem 1. *If a continuous function* f (x) *assumes values of opposite sign at the endpoints of an interval* $[\alpha, \beta]$, *i. e.*, $f(\alpha)\, f(\beta) < 0$, *then the interval will contain at least one root of the equation* $f(x)=0$; *in other words, there will be at least one number* $\xi \in (\alpha, \beta)$[1] *such that* $f(\xi)=0$ (Fig. 9).

The root ξ will definitely be unique if the derivative $f'(x)$ exists and preserves sign within the interval (α, β); that is, if $f'(x)>0$ (or $f'(x)<0$) for $\alpha<x<\beta$ (Fig. 10).

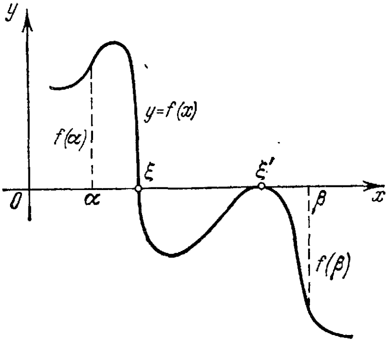


Fig. 9

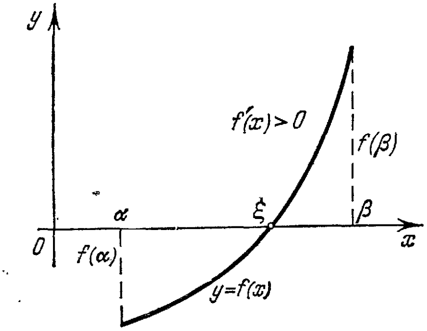


Fig. 10

The process of isolation of roots begins with establishing the signs of the function $f(x)$ at the endpoints $x=a$ and $x=b$ of its domain of existence.

Then the signs of the function $f(x)$ are determined at a number of intermediate points $x=\alpha_1, \alpha_2, \ldots$, the choice of which takes into account the peculiarities of the function $f(x)$. If it turns out that $f(\alpha_k)f(\alpha_{k+1})<0$, then, by virtue of Theorem 1, there is a root of the equation $f(x)=0$ in the interval (α_k, α_{k+1}). We must make sure that this root is the only one. Practically speaking, it is often sufficient, for isolation of the roots, to perform a **halving process**, approximately dividing the given interval (α, β) into two, four, eight, etc., equal parts (up to a certain interval) and to determine the signs of $f(x)$ at the points of division. It is well to recall that an nth-degree algebraic equation

$$a_0x^n+a_1x^{n-1}+\ldots+a_n=0 \qquad (a_0\neq 0)$$

has at most n real roots. Therefore, if for such an equation we obtain $n+1$ sign-changes, then all the roots of the equation have been isolated.

Example 1. Isolate the roots of the equation

$$f(x)\equiv x^3-6x+2=0 \tag{2}$$

---

[1] The notation $\xi \in (\alpha, \beta)$ signifies that the point ξ belongs to the interval (α, β).

Solution. Tabulate the findings roughly as follows:

x	$f(x)$	x	$f(x)$
$-\infty$	$-$	1	$-$
-3	$-$	3	$+$
-1	$+$	$+\infty$	$+$
0	$+$		

Thus, equation (2) has three real roots within the intervals $(-3, -1)$, $(0, 1)$ and $(1, 3)$.

If there is a continuous derivative $f'(x)$ and the roots of the equation

$$f'(x) = 0$$

can readily be computed, the process of isolation of roots of equation (1) may be regularized. Clearly, it is then sufficient to count only the signs of the function $f(x)$ at the points of the zeros of its derivative and at the end points $x = a$ and $x = b$.

Example 2. Isolate the roots of the equation

$$f(x) \equiv x^4 - 4x - 1 = 0 \tag{3}$$

Solution. Here $f'(x) = 4(x^3 - 1)$ and so $f'(x) = 0$ for $x = 1$. We have $f(-\infty) > 0\,(+)$; $f(1) < 0\,(-)$; $f(+\infty) > 0\,(+)$. Consequently, equation (3) has only two real roots, one of which lies in the interval $(-\infty, 1)$, and the other lies in the interval $(1, +\infty)$.

Example 3. Determine the number of real roots of the equation

$$f(x) \equiv x + e^x = 0 \tag{4}$$

Solution. Since $f'(x) = 1 + e^x > 0$ and $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, it follows that (4) has only one real root.

We give an estimate of the error of an approximate root.

Theorem 2. *Let* ξ *be an exact root and* $\bar{x}$ *an approximate root of the equation* $f(x) = 0$, *both located in the same interval* $[\alpha, \beta]$, *and* $|f'(x)| \geqslant m_1 > 0$ *for* $\alpha \leqslant x \leqslant \beta$.1)

Then the following estimate holds true:

$$|\bar{x} - \xi| \leqslant \frac{|f(\bar{x})|}{m_1} \tag{5}$$

---

1) For m_1 we can, for instance, take the least value of $|f'(x)|$ when $\alpha \leqslant x \leqslant \beta$.

Proof. Applying the mean-value theorem, we have

$$f(\bar{x})-f(\xi)=(\bar{x}-\xi)\,f'(c)$$

where c is a value intermediate between $\bar{x}$ and ξ, that is $c \in (\alpha, \beta)$. From this, because $f(\xi)=0$ and $|f'(c)| \geqslant m_1$, we get

$$|f(\bar{x})-f(\xi)|=|f(\bar{x})| \geqslant m_1 |\bar{x}-\xi|$$

Hence

$$|\bar{x}-\xi| \leqslant \frac{|f(\bar{x})|}{m_1}$$

Note. Formula (5) may yield rough results and so is not always convenient to use. For this reason, in practical situations it is common to narrow in some way the general interval (α, β) that contains the root ξ and its approximate value $\bar{x}$, and to assume $|\bar{x}-\xi| \leqslant \beta-\alpha$.

Example 4. As an approximate root of the equation $f(x) \equiv x^4 - x - 1 = 0$ we have $\bar{x}=1.22$. Estimate the absolute error in this root.

Solution. We have $f(\bar{x})=2.2153-1.22-1=-0.0047$.

Since for $\bar{\bar{x}}=1.23$ we get

$$f(\bar{\bar{x}})=2.2888-1.23-1=+0.0588$$

the exact root ξ lies in the interval $(1.22, 1.23)$. The derivative $f'(x)=3x^3-1$ increases monotonically, and so its smallest value in the given interval is

$$m_1=3 \cdot 1.22^3-1=3 \cdot 1.816-1=4.448$$

whence, by formula (5), we get

$$|\bar{x}-\xi| \leqslant \frac{0.0047}{4.448} \approx 0.001$$

Note. Occasionally, in practical situations the accuracy of an approximate root $\bar{x}$ is estimated by how well it satisfies the given equation $f(x)=0$; that is, if the number $|f(\bar{x})|$ is small, then $\bar{x}$ is assumed to be a good approximation to the exact root ξ; but if $|f(\bar{x})|$ is great, then $\bar{x}$ is taken to be a rough value of the exact root ξ. As Figs. 11 and 12 show, this approach is erroneous. One should likewise not forget that if the equation $f(x)=0$ is multiplied by an arbitrary number $N \neq 0$, then we obtain an equivalent equation $Nf(x)=0$, and the number $|Nf(\bar{x})|$ may be made arbitrarily large or small by an appropriate choice of the factor N.

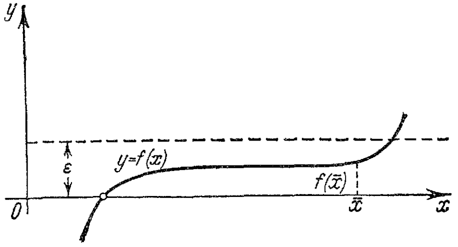


Fig. 11

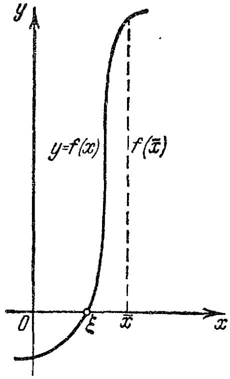


Fig. 12

4.2 GRAPHICAL SOLUTION OF EQUATIONS

The real roots of the equation

$$f(x)=0 \tag{1}$$

can be determined approximately as the abscissas of the points of intersection of the graph of the function $y=f(x)$ with the x-axis (Fig. 9). If equation (1) does not have nearly equal roots, this method can readily be used to isolate them. It is often advisable to replace equation (1) by an equivalent equation (two equations are termed equivalent if they have exactly the same roots):

$$\varphi(x)=\psi(x) \tag{2}$$

where the functions $\varphi(x)$ and $\psi(x)$ are simpler than $f(x)$. Then we construct the graphs of the functions $y=\varphi(x)$ and $y=\psi(x)$, and the desired roots are obtained as the abscissas of the points of intersection of these graphs.

Example 1. Solve the following equation graphically

$$x \log_{10} x = 1 \tag{3}$$

Solution. Write equation (3) as

$$\log_{10} x = \frac{1}{x}$$

The roots of (3) can clearly be found as the abscissas of the points of intersection of the logarithmic curve $y=\log_{10} x$ and the hyperbola $y=\frac{1}{x}$. Constructing these curves (Fig. 13) on coordinate paper, we get an approximate value of the sole root $\xi \approx 2.5$ of equation (3).

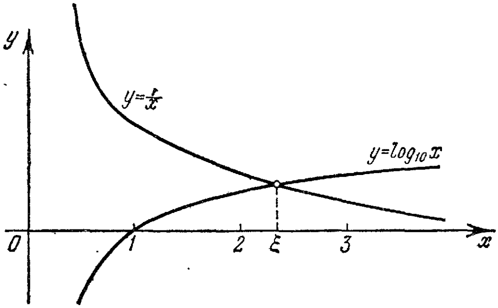


Fig. 13

Finding the roots of equation (2) is simplified if one of the functions $\varphi(x)$ or $\psi(x)$ is linear, say $\varphi(x)=ax+b$. Then the roots of (2) are found as the abscissas of the points of intersection of the curve $y=\psi(x)$ and the straight line $y=ax+b$. This device is particularly advantageous when solving a series of equations of the same type which differ solely in the coefficients a and b of a linear function. Here the graphical construction reduces to finding the points of intersection of a fixed graph, $y=\psi(x)$, and various straight lines. This type obviously includes the three-term equations

$$x^n+ax+b=0$$

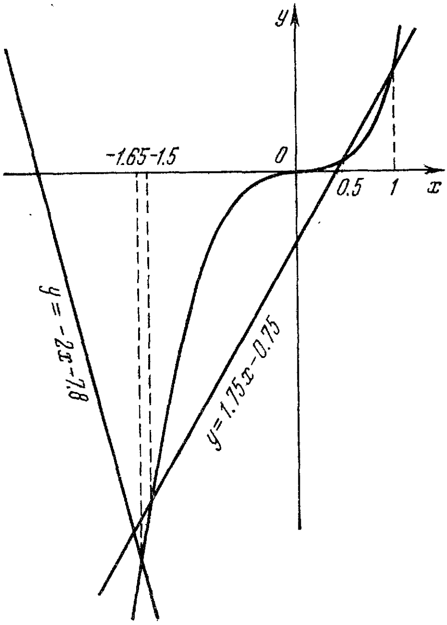


Fig. 14

Example 2. Solve the cubic equations

$$x^3-1.75x+0.75=0$$

and

$$x^3+2x+7.8=0$$

Solution. Construct the cubic parabola $y=x^3$. The desired roots are found as the abscissas of the points of intersection of this parabola by the straight lines (Fig. 14) $y = 1.75x - 0.75$ and $y = -2x - 7.8$. It is clear from the figure that the first equation has three real roots: $x_1 = -1.5$, $x_2 = 0.5$, $x_3 = 1$, and the second equation has only one real root, $x_1 = -1.65$.

It should be noted that although graphical methods of solving equations are very convenient and comparatively simple, they are as a rule used only for a rough determination of the roots. With respect to loss of accuracy, a particularly unfavourable case is when the lines intersect at a very acute angle and practically merge along a certain arc.

One variety of the graphical methods of solving equations are *nomographic methods*, which the reader will find in specialized manuals.

4.3 THE HALVING METHOD

Suppose we have an equation

$$f(x) = 0 \tag{1}$$

where the function $f(x)$ is continuous on $[a, b]$ and $f(a)\, f(b) < 0$.

In order to find a root of (1) lying in the interval $[a, b]$, divide the interval in half. If $f\left(\frac{a+b}{2}\right) = 0$, then $\xi = \frac{a+b}{2}$ is a root of the equation. If $f\left(\frac{a+b}{2}\right) \neq 0$, then we choose that half, $\left[a, \frac{a+b}{2}\right]$ or $\left[\frac{a+b}{2}, b\right]$, at the endpoints of which the function $f(x)$ has opposite signs. The newly reduced interval $[a_1, b_1]$ is again halved and the same investigation is made, etc. Finally, at some stage in the process we get either the exact root of (1) or an infinite sequence of nested intervals $[a_1, b_1]$, $[a_2, b_2]$, ..., $[a_n, b_n]$, ... such that

$$f(a_n)\, f(b_n) < 0 \qquad (n = 1, 2, \ldots) \tag{2}$$

and

$$b_n - a_n = \frac{1}{2^n}(b - a) \tag{3}$$

Since the left endpoints $a_1, a_2, \ldots, a_n, \ldots$ form a monotonic nondecreasing bounded sequence, and the right endpoints $b_1, b_2, \ldots, b_n, \ldots$ form a monotonic nonincreasing bounded sequence, then by (3) there exists a common limit

$$\xi = \lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n$$

Passing to the limit in inequality (2) as $n \to \infty$, we get, by virtue of the continuity of the function $f(x)$, $[f(\xi)]^2 \leqslant 0$, whence $f(\xi) = 0$, which means that ξ is a root of equation (1); it is obvious that

$$0 \leqslant \xi - a_n \leqslant \frac{1}{2^n}(b-a) \tag{4}$$

If the roots of equation (1) are not isolated in the interval $[a, b]$, then this device can be used to find one of the roots of (1).

The method of halving is conveniently used in rough approximations of the root of the given equation since the amount of computation increases substantially with increase in accuracy.

The halving method, by the way, is well suited to electronic computers. The computational routine is set up so that the machine finds the value of the right member of equation (1) at the midpoint of each of the intervals $[a_n, b_n]$ $(n = 1, 2, \ldots)$ and chooses the appropriate half.

Example. Use the halving method to improve a root of the equation

$$f(x) \equiv x^4 + 2x^3 - x - 1 = 0$$

lying in the interval $[0, 1]$.

Solution. We have successively

$$\begin{aligned}
f(0) &= -1, \ f(1) = 1,\\
f(0.5) &= 0.06 + 0.25 - 0.5 - 1 = -1.19,\\
f(0.75) &= 0.32 + 0.84 - 0.75 - 1 = -0.59,\\
f(0.875) &= 0.59 + 1.34 - 0.88 - 1 = +0.05,\\
f(0.8125) &= 0.436 + 1.072 - 0.812 - 1 = -0.304,\\
f(0.8438) &= 0.507 + 1.202 - 0.844 - 1 = -0.135,\\
f(0.8594) &= 0.546 + 1.270 - 0.859 - 1 = -0.043, \text{ etc.}
\end{aligned}$$

We can take

$$\xi \doteq \frac{1}{2}(0.859 + 0.875) = 0.867$$

4.4 THE METHOD OF PROPORTIONAL PARTS (METHOD OF CHORDS)

Using the assumptions of Sec. 4.3, we give a faster method for finding a root ξ of the equation $f(x) = 0$ lying in a specified interval $[a, b]$ such that $f(a)f(b) < 0$.

For the sake of definiteness, suppose $f(a) < 0$ and $f(b) > 0$. Then instead of halving the interval $[a, b]$ it is more natural to divide it in the ratio $-f(a):f(b)$. This yields an approximate value of the root

$$x_1 = a + h_1 \tag{1}$$

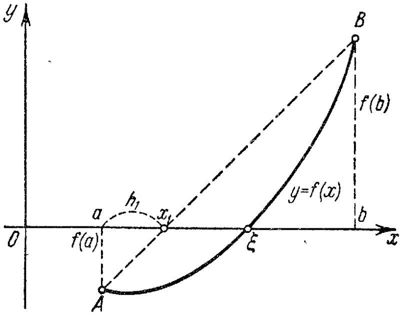
where

$$h_1 = \frac{-f(a)}{-f(a)+f(b)}(b-a) = -\frac{f(a)}{f(b)-f(a)}(b-a) \tag{2}$$

Then, applying this device to the interval $[a, x_1]$ or $[x_1, b]$ at the endpoints of which the function $f(x)$ has opposite signs, we get a second approximation to the root x_2, etc.

Geometrically, the method of proportional parts is equivalent to replacing the curve $y=f(x)$ by a chord that passes through the points $A[a, f(a)]$ and $B[b, f(b)]$ (Fig. 15). Indeed, the equation of the chord AB is

Fig. 15



$$\frac{x-a}{b-a} = \frac{y-f(a)}{f(b)-f(a)}$$

Whence, assuming $x=x_1$ and $y=0$, we get

$$x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a) \tag{1'}$$

Formula (1′) is completely equivalent to formulas (1) and (2).

To prove the convergence of the process, let us assume that the root is isolated and the second derivative $f''(x)$ has a constant sign on the interval $[a, b]$.

Suppose, for definiteness, that $f''(x) > 0$ for $a \leqslant x \leqslant b$ (the case $f''(x) < 0$ reduces to our case if we write the equation as $-f(x)=0$). Then the curve $y=f(x)$ will be convex down and, hence, will be located below its chord AB. Two cases are possible: (1) $f(a) > 0$ (Fig. 16) and (2) $f(a) < 0$ (Fig. 17).

In the former case, the endpoint a is fixed and the successive approximations: $x_0=b$,

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n)-f(a)}(x_n-a) \qquad (n=0, 1, 2, \ldots) \tag{3}$$

form a bounded decreasing monotonic sequence, and

$$a < \xi < \ldots < x_{n+1} < x_n < \ldots < x_1 < x_0$$

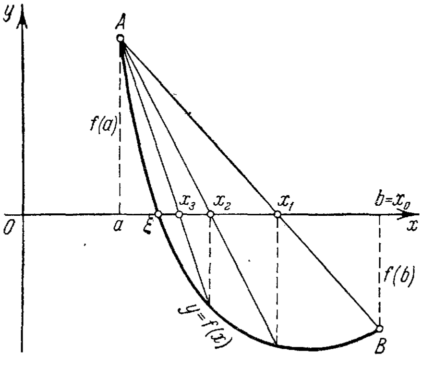


Fig. 16

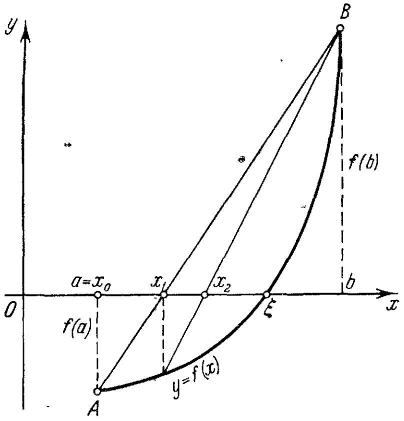


Fig. 17

In the latter case, the endpoint b is fixed and the successive approximations: $x_0 = a$,

$$x_{n+1} = x_n - \frac{f(x_n)}{f(b) - f(x_n)}(b - x_n) \tag{4}$$

form a bounded increasing monotonic sequence, and

$$x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} < \ldots < \xi < b$$

Summarizing, we conclude that: (1) the fixed endpoint is that one for which the sign of the function $f(x)$ coincides with the sign of its second derivative $f''(x)$; (2) the successive approximations x_n lie on the side of the root ξ where the sign of the function $f(x)$ is opposite to the sign of its second derivative $f''(x)$. In both cases, each successive approximation x_{n+1} is closer to the root ξ than the preceding one, x_n. Suppose

$$\bar{\xi} = \lim_{n \to \infty} x_n \qquad (a < \bar{\xi} < b)$$

(the limit exists since the sequence $\{x_n\}$ is bounded and monotonic). Passing to the limit in (3), we have for the first case

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f(\bar{\xi}) - f(a)}(\bar{\xi} - a)$$

whence $f(\bar{\xi}) = 0$. Since it is given that the equation $f(x) = 0$ has only one root ξ on the interval (a, b), it follows that $\bar{\xi} = \xi$, which completes the proof.

By means of the very same limit process in (4), it can be proved that $\bar{\xi} = \xi$ for the second case.

For an estimate of the accuracy of the approximation, we can use formula (5) of Sec. 4.1

$$|x_n - \xi| \leqslant \frac{|f(x_n)|}{m_1}$$

where $|f'(x)| \geqslant m_1$ for $a \leqslant x \leqslant b$.

We give another formula which permits estimating the absolute error in an approximate value x_n if two successive values x_{n-1} and x_n are known.

We will assume that the derivative $f'(x)$ is continuous on the interval $[a, b]$ containing all the approximations and preserves sign; note that

$$0 < m_1 \leqslant |f'(x)| \leqslant M_1 < +\infty \tag{5}$$

For the sake of definiteness, let us assume that the successive approximations x_n to the exact root ξ are generated by formula (3) [a consideration of formula (4) is similar]:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f(x_{n-1}) - f(a)} (x_{n-1} - a)$$

($n = 1, 2, \ldots$) where the endpoint a is fixed. Then, taking note of the fact that $f(\xi) = 0$, we get

$$f(\xi) - f(x_{n-1}) = \frac{f(x_{n-1}) - f(a)}{x_{n-1} - a} (x_n - x_{n-1})$$

Using the mean-value theorem, we get

$$(\xi - x_{n-1}) f'(\xi_{n-1}) = (x_n - x_{n-1}) f'(\bar{x}_{n-1})$$

where $\xi_{n-1} \in (x_{n-1}, \xi)$ and $\bar{x}_{n-1} \in (a, x_{n-1})$. Hence

$$|\xi - x_n| = \frac{|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})|}{|f'(\xi_{n-1})|} |x_n - x_{n-1}| \tag{6}$$

Since $f'(x)$ has constant sign over the interval $[a, b]$ and $\bar{x}_{n-1} \in [a, b]$ and $\xi_{n-1} \in [a, b]$, we plainly obtain

$$|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})| \leqslant M_1 - m_1$$

We therefore derive from formula (6)

$$|\xi - x_n| \leqslant \frac{M_1 - m_1}{m_1} |x_n - x_{n-1}| \tag{7}$$

where for m_1 and M_1 we can take, respectively, the smallest and largest values of the modulus of the derivative $f'(x)$ on the interval $[a, b]$. If the interval $[a, b]$ is so narrow that the inequality

$$M_1 \leqslant 2m_1$$

holds, then from formula (7) we obtain

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|$$

Thus, in this case, as soon as it is clear that

$$|x_n - x_{n-1}| < \varepsilon$$

where ε is the specified limiting absolute error, then it is guaranteed that

$$|\xi - x_n| < \varepsilon$$

Example. Find a positive root of the equation

$$f(x) \equiv x^3 - 0.2x^2 - 0.2x - 1.2 = 0$$

to an accuracy of 0.002.

Solution. First of all isolate the root. Since

$$f(1) = -0.6 < 0 \quad \text{and} \quad f(2) = 5.6 > 0$$

the desired root ξ lies in the interval (1, 2). The resulting interval is great and so we halve it. Since

$$f(1.5) = 1.425$$

it follows that $1 < \xi < 1.5$.

Applying formulas (1) and (2) in succession, we get

$$x_1 = 1 + \frac{0.6}{1.425 + 0.6}(1.5 - 1) = 1 + 0.15 = 1.15,$$

$$f(x_1) = -0.173,$$

$$x_2 = 1.15 + \frac{0.173}{1.425 + 0.173}(1.5 - 1.15) = 1.15 + 0.040 = 1.190,$$

$$f(x_2) = -0.036,$$

$$x_3 = 1.190 + \frac{0.036}{1.425 + 0.036}(1.5 - 1.190) = 1.190 + 0.008 = 1.198,$$

$$f(x_3) = -0.0072$$

Since $f'(x) = 3x^2 - 0.4x - 0.2$ and for $x_3 < x < 1.5$ we have

$$f'(x) \geqslant 3 \cdot 1.198^2 - 0.4 \cdot 1.5 - 0.2 = 3 \cdot 1.43 - 0.8 = 3.49$$

we can take it that

$$0 < \xi - x_3 < \frac{0.0072}{3.49} \approx 0.002$$

Thus, $\xi = 1.198 + 0.002\theta$ where $0 < \theta \leqslant 1$. Note that the exact root of equation (5) is $\xi = 1.2$.

4.5 NEWTON'S METHOD (METHOD OF TANGENTS)

Suppose the root ξ of the equation

$$f(x)=0 \tag{1}$$

is isolated on the interval $[a, b]$ and $f'(x)$ and $f''(x)$ are continuous and preserve signs for $a \leqslant x \leqslant b$. Having found an nth approximation of the root, $x_n \approx \xi \, (a \leqslant x_n \leqslant b)$, we can improve it by *Newton's method* in the following manner.

Set

$$\xi = x_n + h_n \tag{2}$$

where h_n is a small quantity. Then, applying Taylor's formula, we have

$$0 = f(x_n + h_n) \approx f(x_n) + h_n f'(x_n)$$

Consequently,

$$h_n = -\frac{f(x_n)}{f'(x_n)}$$

Inserting this correction into formula (2), we get the next approximation of the root:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (n = 0, 1, 2, \ldots) \tag{3}$$

Geometrically, Newton's method is equivalent to replacing a small arc of the curve $y=f(x)$ by a tangent line drawn to a point of the curve. Indeed, suppose, for definiteness, that $f''(x) > 0$ for $a \leqslant x \leqslant b$ and $f(b) > 0$ (Fig. 18).

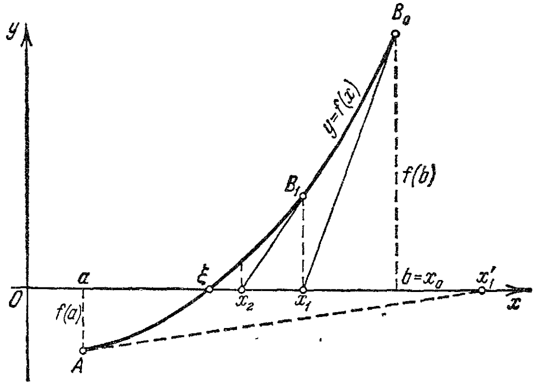


Fig. 18

Choose, say, $x_0 = b$ for which $f(x_0) f''(x_0) > 0$. Draw the tangent line to the curve $y=f(x)$ at the point $B_0[x_0, f(x_0)]$. For the first approximation x_1 of the root ξ let us take the abscissa of the point of intersection of this tangent with the x-axis. Through the point $B_1[x_1, f(x_1)]$ again draw a tangent line whose abscissa of the intersection point with the x-axis yields a second approximation x_2 of the root ξ, and so on (Fig. 18). It is plain that the equation of the tangent at the point $B_n[x_n, f(x_n)]$ $(n=0, 1, 2, \ldots)$ is

$$y-f(x_n)=f'(x_n)(x-x_n)$$

Putting $y=0$, $x=x_{n+1}$, we get formula (3)

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}$$

Note that if in our case we put $x_0=a$ and hence $f(x_0)\,f''(x_0)<0$, then drawing the tangent to the curve $y=f(x)$ at the point $A[a, f(a)]$, we would get point x_1' (Fig. 18) lying outside the interval $[a, b]$; in other words, Newton's method is impractical for such a choice of the initial value. Thus, in the given case, a "good" initial approximation x_0 is one for which the inequality

$$f(x_0)\,f''(x_0)>0 \tag{4}$$

is valid. We will now prove that this rule is general.

Theorem. *If* f (a) f (b) < 0, *and* f′ (x) *and* f″ (x) *are nonzero and preserve signs over* a ⩽ x ⩽ b, *then, proceeding from the initial approximation* $x_0 \in [a, b]$ *which satisfied* (4), *it is possible, by using Newton's method* [*formula* (3)], *to compute the sole root* ξ *of equation* (1) *to any degree of accuracy.*

Proof. For example, suppose $f(a)<0$, $f(b)>0$, $f'(x)>0$, $f''(x)>0$ for $a\leqslant x\leqslant b$ (the other cases are considered similarly). By inequality (4) we have $f(x_0)>0$ (we can, say, take $x_0=b$).

By mathematical induction we prove that all approximations $x_n>\xi$ $(n=0, 1, 2, \ldots)$ and, hence, $f(x_n)>0$. Indeed, first of all, $x_0>\xi$.

Now let $x_n>\xi$. Put

$$\xi=x_n+(\xi-x_n)$$

Using Taylor's formula, we get

$$0=f(\xi)=f(x_n)+f'(x_n)(\xi-x_n)+\frac{1}{2}f''(c_n)(\xi-x_n)^2 \tag{5}$$

where $\xi<c_n<x_n$.

Since $f''(x)>0$, we have

$$f(x_n)+f'(x_n)(\xi-x_n)<0$$

and, hence,

$$x_{n+1}=x_n-\frac{f(x_n)}{f''(x_n)}>\xi$$

which is what we set out to prove.

Taking into consideration the signs of $f(x_n)$ and $f'(x_n)$ we have, from formula (3), $x_{n+1} < x_n$ $(n = 0, 1, \ldots)$, that is to say, the successive approximations $x_0, x_1, \ldots, x_n, \ldots$ form a bounded decreasing monotonic sequence. Therefore, the limit $\bar{\xi} = \lim\limits_{n \to \infty} x_n$ exists.

Passing to the limit in (3), we have

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f'(\bar{\xi})}$$

or $f(\bar{\xi}) = 0$, whence $\bar{\xi} = \xi$, and the proof is complete.

For this reason, when applying Newton's method, one should be guided by the following rule: *for the initial point* x_0 *choose the end of the interval* (a, b) *associated with an ordinate of the same sign as the sign of* $f''(x)$.

Note 1. If: (1) the function $f(x)$ is defined and continuous over $-\infty < x < +\infty$; (2) $f(a)\,f(b) < 0$; (3) $f'(x) \neq 0$ for $a \leqslant x \leqslant b$; (4) $f''(x)$ exists everywhere and preserves sign, then any value $c \in [a, b]$ may be taken for the initial approximation x_0 when using Newton's method to find a root of the equation $f(x) = 0$ lying in the interval (a, b). One can, for instance, take $x_0 = a$ or $x_0 = b$.

Indeed, suppose, say, $f'(x) > 0$ for $a \leqslant x \leqslant b$, $f''(x) > 0$ and $x_0 = c$, where $a \leqslant c \leqslant b$. If $f(c) = 0$, then the root $\xi = c$ and the problem is solved. If $f(c) > 0$, then the foregoing reasoning holds true and the Newton process with initial value c converges to the root $\xi \in (a, b)$.

Finally, if $f(c) < 0$, then we find the value

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = c - \frac{f(c)}{f'(c)} > c$$

Using Taylor's formula, we have

$$f(x_1) = f(c) - \frac{f(c)}{f'(c)} f'(c) + \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) = \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) > 0$$

where $\bar{c}$ is a certain value intermediate between c and x_1.

Thus

$$f(x_1)\,f''(x_1) > 0$$

Besides, from the condition $f''(x) > 0$ it follows that $f'(x)$ is an increasing function and, hence, $f'(x) > f'(a) > 0$ for $x > a$. It is thus possible to take x_1 for the initial value of the Newton process converging to some root $\bar{\xi}$ of the function $f(x)$ such that $\bar{\xi} > c \geqslant a$. Since, because the derivative $f'(x)$ is positive when $x > a$, the function $f(x)$ has a unique root in the interval $(a, +\infty)$, it follows that

$$\bar{\xi} = \xi \in (a, b)$$

A similar argument can be devised for other combinations of signs of the derivatives $f'(x)$ and $f''(x)$.

Note 2. From formula (3) it is clear that the larger the numerical value of the derivative $f'(x)$ in the neighbourhood of the given root, the smaller the correction that has to be added to the nth approximation in order to obtain the $(n+1)$th approximation. Newton's method is therefore particularly convenient when the graph of the function is steep in the neighbourhood of the given root. But if the numerical value of the derivative $f'(x)$ is small near the root, then the corrections will be great, and computing the root by this method may prove to be an extended procedure or sometimes even impossible. To summarize, then: do not use Newton's method to solve an equation $f(x)=0$ if the curve $y=f(x)$ near the point of intersection with the x-axis is nearly horizontal.

To estimate the error of the nth approximation x_n, one can use the general formula (5) of Sec. 4.1:

$$|\xi-x_n|\leqslant\frac{|f(x_n)|}{m_1} \tag{6}$$

where m_1 is the smallest value of $|f'(x)|$ in the interval $[a, b]$.

We now derive another formula for estimating the accuracy of the approximation x_n. Applying Taylor's formula, we have

$$\begin{aligned} f(x_n)&=f[x_{n-1}+(x_n-x_{n-1})]=\\ &=f(x_{n-1})+f'(x_{n-1})(x_n-x_{n-1})+\frac{1}{2}f''(\xi_{n-1})(x_n-x_{n-1})^2 \end{aligned} \tag{7}$$

where $\xi_{n-1}\in(x_{n-1}, x_n)$. Since, by virtue of the definition of the approximation x_n, we have

$$f(x_{n-1})+f'(x_{n-1})(x_n-x_{n-1})=0$$

it follows, from (7), that

$$|f(x_n)|\leqslant\frac{1}{2}M_2(x_n-x_{n-1})^2$$

where M_2 is the largest value of $|f''(x)|$ on the interval $[a, b]$. Consequently, on the basis of formula (6) we finally get

$$|\xi-x_n|\leqslant\frac{M_2}{2m_1}(x_n-x_{n-1})^2 \tag{8}$$

If the Newton process converges, then $x_n-x_{n-1}\to 0$ as $n\to\infty$. And so for $n\geqslant N$ we have

$$|\xi-x_n|\leqslant|x_n-x_{n-1}|$$

that is, the "stabilized" initial decimal places of the approximations x_{n-1} and x_n are correct beginning with a certain approximation.

Note that in the general case a coincidence, up to ε, of two successive approximations x_{n-1} and x_n does not in the least guarantee that the value of x_n and the exact root ξ (Fig. 19) coincide to within the same degree of accuracy.

We will also derive a formula that connects the absolute errors in two successive approximations x_n and x_{n+1}. From (5) we have

$$\xi = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2} \cdot \frac{f''(c_n)}{f'(x_n)} (\xi - x_n)^2$$

where $c_n \in (x_n, \xi)$, whence, taking into account (3), we have

$$\xi - x_{n+1} = -\frac{1}{2} \cdot \frac{f''(c_n)}{f'(x_n)} (\xi - x_n)^2$$

and, consequently,

$$|\xi - x_{n+1}| \leqslant \frac{M_2}{2m_1} (\xi - x_n)^2 \qquad (9)$$

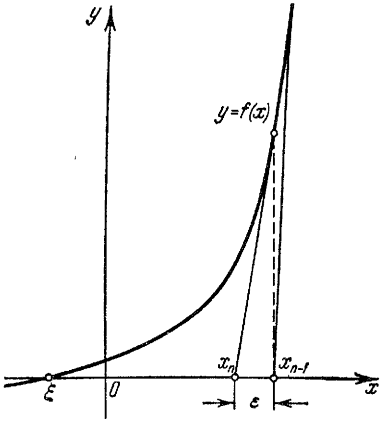


Fig. 19

Formula (9) ensures a rapid convergence of the Newton process if the initial approximation x_0 is such that

$$\frac{M_2}{2m_1} |\xi - x_0| \leqslant q < 1$$

In particular, if

$$\mu = \frac{M_2}{2m_1} \leqslant 1 \quad \text{and} \quad |\xi - x_n| < 10^{-m}$$

then from (9) we get

$$|\xi - x_{n+1}| < 10^{-2m}$$

That is, in this case if the approximation x_n is correct to m decimal places, the next approximation x_{n+1} will be correct to at least $2m$ places; in other words, if $\mu \leqslant 1$, then Newton's method ensures a doubling of correct decimal places of the desired root ξ at every step.

Example 1. Using Newton's method, compute a negative root of the equation $f(x) \equiv x^4 - 3x^2 + 75x - 10{,}000 = 0$ correct to five places.

Solution. Setting $x = 0, -10, -100, \ldots$, in the left member, we get $f(0) = -10{,}000$, $f(-10) = -1050$, $f(-100) \approx +10^8$.

Thus, the desired root ξ lies in the interval $-100 < \xi < -10$. Reduce the interval found. Since $f(-11) = 3453$, then $-11 < \xi < -10$. In this latter interval, $f'(x) < 0$ and $f''(x) > 0$. Since $f(-11) > 0$ and $f''(-11) > 0$, we can take $x_0 = -11$ for the initial approximation. The successive approximations x_n $(n=1, 2, \ldots)$ are computed in accord with the following scheme:

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n=-\frac{f(x_n)}{f'(x_n)}$
0	-11	3453	-5183	0.7
1	-10.3	134.3	-4234	0.03
2	-10.27	37.8	-4196	0.009
3	-10.261	0.2	—	—

Stopping with $n=3$, verify the sign of the value $f(x_n+0.001)=$ $=f(-10.260)$. Since $f(-10.260)<0$, it follows that $-10.261<$ $<\xi<-10.260$ and either of these numbers yields the required approximation.

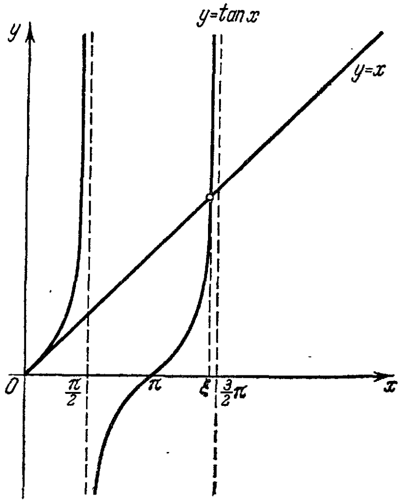


Fig. 20

Example 2. Use Newton's method to find the smallest positive root of the equation $\tan x = x$ to within 0.0001.

Solution. Plotting the graphs of the curves $y=\tan x$ and $y=x$ (Fig. 20), we conclude that the desired root ξ lies in the interval $\pi<\xi<\frac{3\pi}{2}$. Rewriting the equation in the form

$$f(x) \equiv \sin x - x\cos x = 0$$

we have

$$f'(x) = x\sin x,$$
$$f''(x) = \sin x + x\cos x$$

Whence $f'(x)<0$ and $f''(x)<0$ for $\pi<x<\frac{3\pi}{2}$. Since $f\left(\frac{3\pi}{2}\right)=-1$, for the initial approximation we can take $x_0=\frac{3\pi}{2}$. Perform the computations according to the following scheme.

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n=-\frac{f(x_n)}{f'(x_n)}$
0	$\frac{3\pi}{2}=4.71239$ (270°)	-1	-4.712	-0.212 $(\approx -12°10')$
1	4.50004 (257°50′)	-0.0291	-4.399	-0.0066 $(\approx -22'44'')$
2	4.49343 (257°27′16″)	-0.00003	—	—

In estimating the error in the approximate value x_n, note that the successive approximations x_n $(n=0, 1, 2, \ldots)$ monotonically decrease due to the negativity of the second derivative $f''(x)$, and $f(x_n) < 0$. For this reason, we can take $\bar{x}_n < \xi < x_n$, where $\bar{x}_n$ is a value from the interval $\left(\pi, \frac{3\pi}{2}\right)$ such that $f(\bar{x}_n) > 0$. The value of $\bar{x}_n$ can easily be found by inspection. [True, one could take $\bar{x}_n = \pi$, but this is not advantageous because $f'(\pi) = 0$.] Thus, for instance, for $n=2$ and assuming approximately

$$\bar{x}_2 = 4.49340 = \text{arc } 257°27'12''$$

we have

$$\begin{aligned} f(\bar{x}_2) &= \sin 257°27'12'' - 4.49340 \cdot \cos 257°27'12'' = \\ &= -0.97612 + 4.49340 \cdot 0.21724 = \\ &= -0.97612 + 0.97614 = +0.00002 \end{aligned}$$

Hence $\bar{x}_2$ is chosen correctly and

$$4.49340 < \xi < 4.49343$$

We can put

$$\xi = 4.4934$$

which is correct to all digits written.

The estimate of the error in the value of x_2 can readily be made more exact. Since, for $x \in [\bar{x}_2, x_2]$, the derivative $f'(x)$ decreases and $f'(x) < 0$, then

$$m_1 = \min |f'(x)| = |f'(\bar{x}_2)|$$

whence

$$m_1 = 4.49340 \cdot 0.97612 > 4$$

and, consequently,

$$|\xi - x_2| \leqslant \frac{|f(x_2)|}{4} = \frac{0.00003}{4} < 10^{-5}$$

Thus

$$\xi = 4.49343 - 0.00001\theta$$

where $0 < \theta < 1$.

Example 3. Consider the equation

$$f(x) = 0 \tag{10}$$

where $f''(x)$ is continuous and preserves sign over $-\infty < x < +\infty$. By Rolle's theorem, equation (10) cannot have more than two real roots. Let us note two cases of practical interest.

I. Suppose

$$f(x_0) f'(x_0) < 0, \quad f(x_0) f''(x) < 0$$

(Fig. 21).

Then (10) has the unique root ξ in the interval (x_0, x_1), where

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

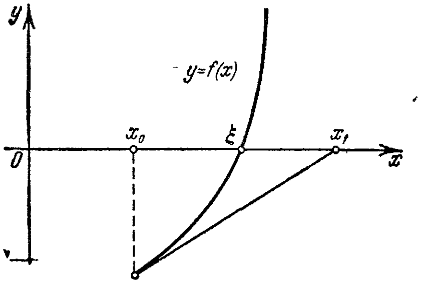

y
y=f(x)
x_0
ξ
x_1
0
x

Fig. 21

The root ξ may be computed to the given accuracy by Newton's method.

II. Let

$$f'(x_0) = 0, \qquad f(x_0) f''(x) < 0$$

Then equation (10) has two roots ξ and ξ' in the interval $(-\infty, +\infty)$ (Fig. 22).

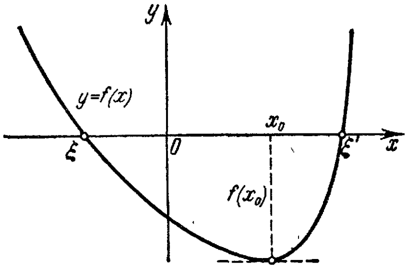

y
y=f(x)
x_0
ξ
0
ξ'
x
$f(x_0)$

Fig. 22

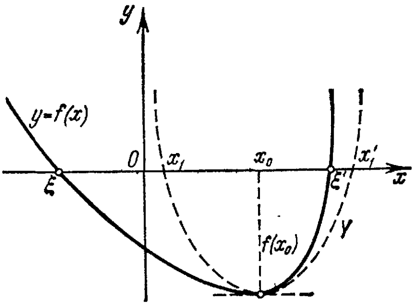

y
y=f(x)
0
x_1
x_0
x_1'
x
ξ
ξ'
Y
$f(x_0)$

Fig. 23

Transforming the left member of (10) by Taylor's formula, we have approximately

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 = 0$$

or

$$f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 = 0$$

whence for the roots ξ and ξ' we get the initial approximations

$$x_1 = x_0 - \sqrt{-\frac{2f(x_0)}{f''(x_0)}}$$

and

$$x_1' = x_0 + \sqrt{-\frac{2f(x_0)}{f''(x_0)}}$$

which are the abscissas of the points of intersection of the parabola

$$Y = f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

with the x-axis (Fig. 23). Subsequent improvements in the roots can be carried out by the usual Newton method.

Assertions I and II are geometrically obvious. It is left to the reader to carry out a rigorous proof.

4.6 MODIFIED NEWTON METHOD

If the derivative $f'(x)$ varies but slightly on the interval $[a, b]$, then in formula (3) of the preceding section we can put

$$f'(x_n) \approx f'(x_0) \tag{1}$$

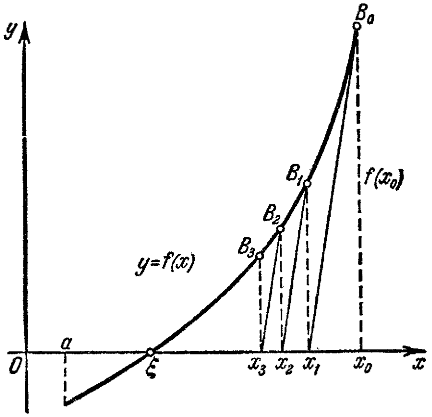


Fig. 24

From this, for the root ξ of the equation $f(x) = 0$ we get the successive approximations

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \qquad (n = 0, 1, \ldots) \tag{2}$$

Geometrically, this method signifies that we replace the tangents at the points $B_n[x_n, f(x_n)]$ by straight lines parallel to the tangent to the curve $y = f(x)$ at its fixed point $B_0[x_0, f(x_0)]$ (Fig. 24).

Formula (1) relieves us of the necessity to compute the values of the derivative $f'(x_n)$ each time; and so this formula is very useful if $f'(x_n)$ is complicated. It can be proved that on the assumption of constancy of signs of the derivatives $f'(x)$ and $f''(x)$ the successive approximations (2) yield a convergent process.

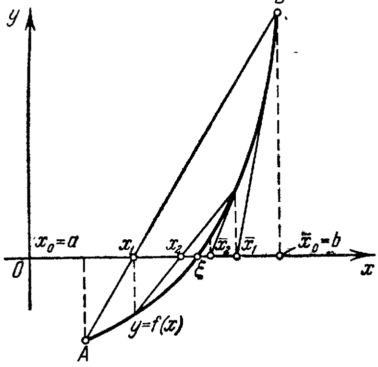
4.7 COMBINATION METHOD

Suppose $f(a) \cdot f(b) < 0$ and $f'(x)$ and $f''(x)$ preserve signs on the interval $[a, b]$. Combining the method of proportional parts and Newton's method, we obtain a method, at each stage of which we find minor (too small) and major (too large) approximations to the exact root ξ of the equation $f(x) = 0$.

A consequence of this is that the digits common to x_n and $\bar{x}_n$ must definitely belong to the exact root ξ. There are four theoretically possible cases here:

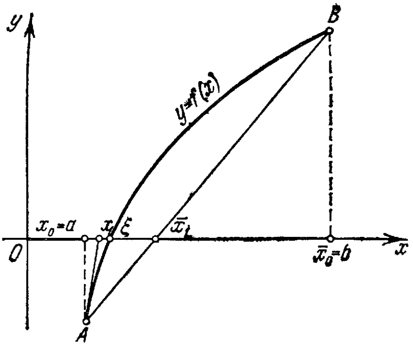
(1) $f'(x) > 0$, $f''(x) > 0$ (Fig. 25),

(2) $f'(x) > 0$, $f''(x) < 0$ (Fig. 26),



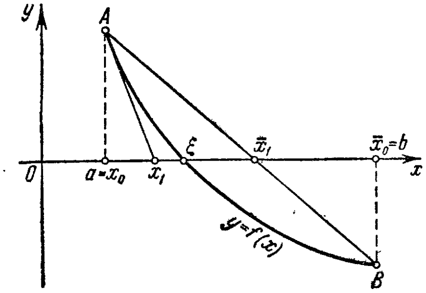
y
B
$x_0=a$
x_1
x_2
$\bar{x}_2$
$\bar{x}_1$
$\bar{x}_0=b$
0
x
ξ
$y=f(x)$
A

Fig. 25



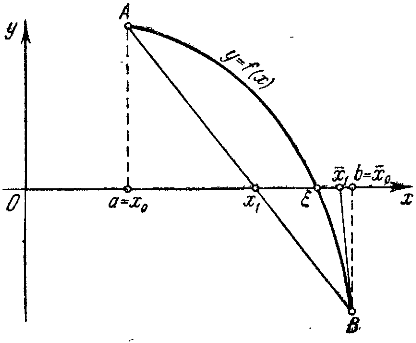
y
B
$y=f(x)$
$x_0=a$
x_1
ξ
$\bar{x}_1$
0
$\bar{x}_0=b$
x
A

Fig. 26



y
A
ξ
$\bar{x}_1$
$\bar{x}_0=b$
0
$a=x_0$
x_1
x
$y=f(x)$
B

Fig. 27



A
y
$y=f(x)$
$\bar{x}_1$
$b=\bar{x}_0$
0
$a=x_0$
x_1
ξ
x
B

Fig. 28

(3) $f'(x)<0$, $f''(x)>0$ (Fig. 27),
(4) $f'(x)<0$, $f''(x)<0$ (Fig. 28).

We confine our analysis to the first case. The remaining cases are studied in similar fashion and the character of the computations is easy to understand on the basis of the figures. It is worth noting that these cases can be reduced to the first one if we replace the equation $f(x)=0$ by the equivalent equations $-f(x)=0$ or $\pm f(-z)=0$, where $z=-x$.

Thus, suppose $f'(x)>0$ and $f''(x)>0$ for $a \leqslant x \leqslant b$. Put $x_0=a$, $\bar{x}_0=b$ and

$$x_{n+1}=x_n-\frac{f(x_n)}{f(\bar{x}_n)-f(x_n)}(\bar{x}_n-x_n), \tag{1}$$

$$\bar{x}_{n+1}=\bar{x}_n-\frac{f(\bar{x}_n)}{f'(\bar{x}_n)} \qquad (n=0,\ 1,\ 2,\ \ldots) \tag{1'}$$

(At each step the method of chords is applied to a new interval $[x_n, \bar{x}_n]$.)

From what was proved above (Secs. 4.5 and 4.6) it follows that

$$x_n<\xi<\bar{x}_n$$

and

$$0<\xi-x_n<\bar{x}_n-x_n \tag{2}$$

If the permissible absolute error in an approximate root x_n is specified beforehand and is equal to ε, then the process of approach terminates as soon as we see that $\bar{x}_n-x_n<\varepsilon$. At the termination of the process, it is best, for the value of the root ξ, to take the arithmetic mean of the last values obtained:

$$\bar{\xi}=\frac{1}{2}(x_n+\bar{x}_n)$$

Example. Compute to within 0.0005 the sole positive root of the equation

$$f(x)\equiv x^5-x-0.2=0$$

Solution. Since $f(1)<0$ and $f(1.1)>0$, the root lies in the interval (1, 1.1). We have

$$f'(x)=5x^4-1 \text{ and } f''(x)=20x^3$$

In the chosen interval, $f'(x)>0$, $f''(x)>0$, which means the signs of the derivatives are preserved.

Let us apply the combination method assuming $x_0=1$ and $\bar{x}_0=1.1$. Since

$$f(x_0)=f(1)=-0.2, \qquad f(\bar{x}_0)=f(1.1)=0.3105,$$
$$f'(\bar{x}_0)=f'(1.1)=6.3205$$

the formulas (1) and (1′) yield

$$x_1 = 1 + \frac{0.1 \cdot 0.2}{0.51051} \approx 1.039, \qquad \bar{x}_1 = 1.1 - \frac{0.31051}{6.3205} \approx 1.051$$

Since $\bar{x}_1 - x_1 = 0.012$, the accuracy is not sufficient. We find the next pair of approximations:

$$x_2 = 1.039 + \frac{0.012 \cdot 0.0282}{0.0595} \approx 1.04469, \qquad \bar{x}_2 = 1.051 - \frac{0.0313}{5.1005} \approx 1.04487$$

Here, $\bar{x}_2 - x_2 = 0.00018$, which shows the desired degree of accuracy has been achieved. We can put

$$\bar{\xi} = \frac{1}{2}(1.04469 + 1.04487) = 1.04478 \approx 1.045$$

with absolute error less than

$$\frac{1}{2} \cdot 0.00018 + 0.00022 = 0.00031 < \frac{1}{2} \cdot 10^{-3}$$

4.8 THE METHOD OF ITERATION

One of the most important methods in the numerical solution of equations is the *method of iteration* (often also called the *method of successive approximations*). Essentially, this method consists in the following. Suppose we have an equation

$$f(x) = 0 \tag{1}$$

where $f(x)$ is a continuous function and it is required to determine its real roots. Replace (1) with an equivalent equation,

$$x = \varphi(x) \tag{2}$$

In some way choose a roughly approximate value of the root, x_0, and substitute it into the right member of (2) to get a number

$$x_1 = \varphi(x_0) \tag{3}$$

Now inserting x_1 in the right member of (3) in place of x_0, we get a new number $x_2 = \varphi(x_1)$. Repeating this process, we get a sequence of numbers

$$x_n = \varphi(x_{n-1}) \qquad (n = 1, 2, \ldots) \tag{4}$$

If this sequence is convergent, that is, if there exists a limit $\xi = \lim\limits_{n \to \infty} x_n$, then, by passing to the limit in (4) and assuming the function $\varphi(x)$ to be continuous, we find

$$\lim_{n \to \infty} x_n = \varphi\left(\lim_{n \to \infty} x_{n-1}\right)$$

or

$$\xi = \varphi(\xi) \tag{5}$$

Thus, the limit ξ is a root of (2) and can be computed by formula (4) to any desired degree of accuracy.

Geometrically, the method of iteration can be explained as follows. Plot on an xy-plane the graphs of the functions $y=x$ and $y=\varphi(x)$. Each real root ξ of equation (2) is an abscissa of the point of intersection M of the curve $y=\varphi(x)$ with the straight line $y=x$ (Fig. 29).

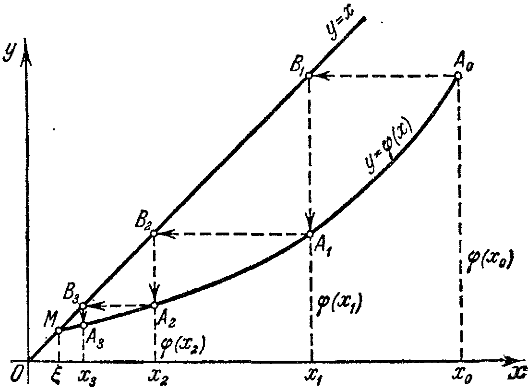


Fig. 29

Starting from a point $A_0[x_0, \varphi(x_0)]$, we construct the polygonal line $A_0B_1A_1B_2A_2\ldots$ ("staircase"), the segments of which are alternately parallel to the x-axis and to the y-axis, the vertices A_0, A_1, A_2, ... lie on the curve $y=\varphi(x)$, and the vertices B_1, B_2, B_3, ... lie on the straight line $y=x$. The common abscissas of the points A_1 and B_1, A_2 and B_2, ..., will obviously be, respectively, the successive approximations x_1, x_2, ... to the root ξ.

It is also possible (Fig. 30) to have a different kind of polygonal

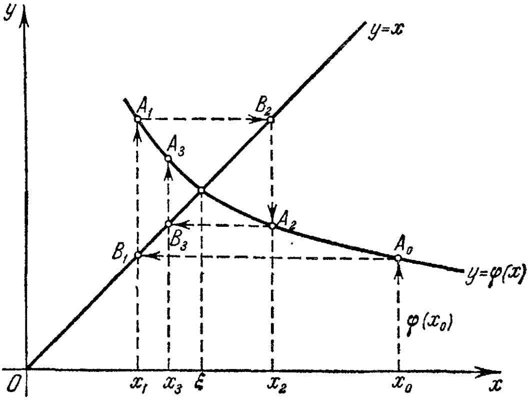


Fig. 30

line $A_0B_1A_1B_2A_2\ldots$ ("spiral"). It is readily seen that the "staircase" solution is obtained if the derivative $\varphi'(x)$ is positive and the "spiral" solution if $\varphi'(x)$ is negative.

In Fig. 29, the curve $y=\varphi(x)$ slopes in the neighbourhood of the root ξ, that is, $|\varphi'(x)|<1$ and the process of iteration converges. However, if we consider the case where $|\varphi'(x)|>1$, then the process of iteration may be divergent (Fig. 31). Hence, to apply the method of iteration practically, we have to ascertain the sufficient conditions for the convergence of the iteration process.

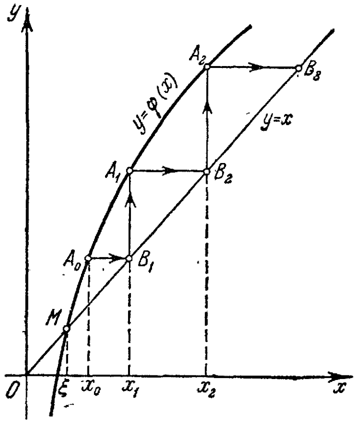


Fig. 31

Theorem 1. *Let a function* φ(x) *be defined and differentiable on an interval* [a, b] *with all values* φ(x) ∈ [a, b].

Then if there exists a proper fraction q[1] *such that*

$$|\varphi'(x)| \leqslant q < 1 \tag{6}$$

for a < x < b, *then:* (1) *the process of iteration*

$$x_n = \varphi(x_{n-1}) \qquad (n=1, 2, \ldots) \tag{7}$$

converges irrespective of the initial value $x_0 \in [a, b]$; (2) *the limiting value*

$$\xi = \lim_{n\to\infty} x_n$$

is the only root of the equation

$$x = \varphi(x) \tag{8}$$

on the interval [a, b].

Proof. We consider two successive approximations

$$x_n = \varphi(x_{n-1}) \text{ and } x_{n+1} = \varphi(x_n)$$

(which are definitely meaningful by virtue of the conditions of the theorem), whence

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1})$$

Applying the mean-value theorem, we have

$$x_{n+1} - x_n = (x_n - x_{n-1})\,\varphi'(\bar{x}_n)$$

---

[1] For the number q we can take the least value or a lower bound of the modulus of the derivative $|\varphi'(x)|$ for $a \leqslant x \leqslant b$.

where $\bar{x}_n \in (x_{n-1}, x_n)$. And so, on the basis of (6), we get

$$|x_{n+1}-x_n| \leqslant q\,|x_n-x_{n-1}| \tag{9}$$

From this, assigning the values $n=1, 2, 3, \ldots$ in succession, we derive

$$\begin{aligned} &|x_2-x_1| \leqslant q\,|x_1-x_0|,\\ &|x_3-x_2| \leqslant q\,|x_2-x_1| \leqslant q^2|x_1-x_0|,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdot\\ &|x_{n+1}-x_n| \leqslant q^n\,|x_1-x_0| \end{aligned} \tag{10}$$

Let us consider the series

$$x_0+(x_1-x_0)+(x_2-x_1)+\ldots+(x_n-x_{n-1})+\ldots \tag{11}$$

for which our successive approximations x_n are $(n+1)$th partial sums, that is

$$x_n=S_{n+1}$$

By the inequality (10), the terms of the series (11) are less in absolute value than the corresponding terms of a geometric progression with ratio $q<1$, and so the series (11) converges and converges absolutely. Hence

$$\lim_{n\to\infty} S_{n+1}=\lim_{n\to\infty} x_n=\xi$$

and obviously $\xi\in[a, b]$.

Passing to the limit in (7), we obtain, by virtue of the continuity of the function $\varphi(x)$,

$$\xi=\varphi(\xi) \tag{12}$$

Thus, ξ is a root of equation (8), which does not have any other root on the interval $[a, b]$. Indeed, if

$$\bar{\xi}=\varphi(\bar{\xi}) \tag{13}$$

then from (12) and (13) we obtain

$$\bar{\xi}-\xi=\varphi(\bar{\xi})-\varphi(\xi)$$

and, consequently,

$$(\bar{\xi}-\xi)\,[1-\varphi'(c)]=0 \tag{14}$$

where $c\in[\xi, \bar{\xi}]$. Since the expression in square brackets in (14) is not zero, it follows that $\xi=\bar{\xi}$, that is, the root ξ is the only one.

Note 1. The theorem remains valid if the function $\varphi(x)$ is defined and differentiable in the infinite interval $-\infty<x<+\infty$, and inequality (6) holds true when $x\in(-\infty, +\infty)$.

Note 2. Under the conditions of Theorem 1, the method of iteration converges *for any choice of the initial value* x_0 in $[a, b]$. For this reason it is *self-correcting*, that is an individual error in the computations that does not go beyond the limits of the interval $[a, b]$ will not affect the final result since an erroneous value may be regarded as a new initial value x_0. Only the amount of work may increase. The property of self-correction makes the method of iteration one of the most reliable computational methods. Quite naturally, systematic errors in applying this method can prevent one from obtaining the result required.

Estimate of an approximation. From formula (10) we have

$$|x_{n+p}-x_n| \leqslant |x_{n+p}-x_{n+p-1}|+|x_{n+p-1}-x_{n+p-2}|+\ldots$$
$$\ldots+|x_{n+1}-x_n| \leqslant q^{n+p-1}|x_1-x_0|+q^{n+p-2}|x_1-x_0|+\ldots$$
$$\ldots+q^n|x_1-x_0|=q^n|x_1-x_0|(1+q+q^2+\ldots+q^{p-1})$$

Summing the geometric progression, we obtain

$$|x_{n+p}-x_n \leqslant | q^n |x_1-x_0| \frac{1-q^p}{1-q} < \frac{q^n}{1-q}|x_1-x_0|$$

Allowing the number p to go to infinity and taking into account that $\lim\limits_{p\to\infty} x_{n+p}=\xi$, we finally get

$$|\xi-x_n| \leqslant \frac{q^n}{1-q}|x_1-x_0| \tag{15}$$

From this it is clear that the convergence of the process of iteration will be the faster, the smaller the number q.

Approximations can be estimated by another formula, which finds application in certain cases. Let

$$f(x)=x-\varphi(x)$$

It is obvious that $f'(x)=1-\varphi'(x) \geqslant 1-q$, whence, noting that $f(\xi)=0$, we get

$$|x_n-\varphi(x_n)|=|f(x_n)-f(\xi)|=|x_n-\xi||f'(\bar{x}_n)| \geqslant (1-q)|x_n-\xi|$$

where $\bar{x}_n \in (x_n, \xi)$ and, hence,

$$|x_n-\xi| \leqslant \frac{|x_n-\varphi(x_n)|}{1-q} \tag{16}$$

that is

$$|\xi-x_n| \leqslant \frac{|x_{n+1}-x_n|}{1-q} \tag{16'}$$

Using formula (9) we also get

$$|\xi-x_n| \leqslant \frac{q}{1-q}|x_n-x_{n-1}| \tag{16''}$$

whence it follows, in particular, that if $q \leqslant \frac{1}{2}$, then

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|$$

In this case, from the inequality $|x_n - x_{n-1}| < \varepsilon$ follows the inequality

$$|\xi - x_n| < \varepsilon$$

Note. There is a widespread opinion that if when using the method of iteration two successive approximations x_{n-1} and x_n coincide to within the specified accuracy ε (for instance, the first m decimals are stabilized in these approximations), then the equality $\xi \approx x_n$ holds true with the same accuracy (that is, in particular, in the given example the approximate number x_n is correct to m places!). As Fig. 32 so vividly reveals, this assertion is erroneous in the general case. What is more, it is easy to demonstrate that if $\varphi'(x)$ is close to unity, then the quantity $|\xi - x_n|$ may be large, although the quantity $|x_n - x_{n-1}|$ is extremely small.

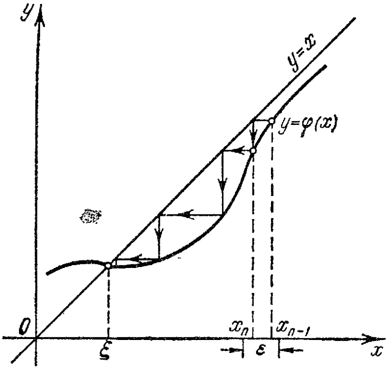


Fig. 32

Formula (16″) enables one to estimate the error in the approximate value x_n from the discrepancy between two successive approximations x_{n-1} and x_n.

The process of iteration should be continued until the inequality

$$|x_n - x_{n-1}| \leqslant \frac{1-q}{q}\varepsilon$$

holds true for the two successive approximations x_{n-1} and x_n; here, ε is the specified limiting absolute error of the root ξ and $|\varphi'(x)| \leqslant q$. Then, by virtue of formula (16″), the inequality

$$|\xi - x_n| \leqslant \varepsilon$$

is valid; that is,

$$\xi = x_n \pm \varepsilon$$

Note that if

$$x_n = \varphi(x_{n-1})$$

and

$$\xi = \varphi(\xi)$$

then

$$|\xi - x_n| = |\varphi(\xi) - \varphi(x_{n-1})| =$$

$$= |\xi - x_{n-1}|\,|\varphi'(\bar{x}_{n-1})| \leqslant q\,|\xi - x_{n-1}| \qquad [\bar{x}_{n-1} \in (x_{n-1}, \xi)]$$

that is,

$$|\xi - x_n| \leqslant |\xi - x_{n-1}|$$

Thus, in a convergent iterative process the error $|\xi - x_n|$ tends to zero monotonically, which is to say, each successive value x_n is more exact than the preceding value x_{n-1}. In all these conclusions we of course **ignore rounding errors**; it is assumed that the successive approximations are found exactly.

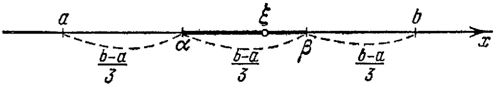
Ordinarily, in practical situations a crude technique is used to establish the existence of a root ξ of equation (2) and then by the method of iteration it is required to obtain a sufficiently exact approximate value of the root; inequality (6) is valid only within a certain neighbourhood (a, b) of this root. If the choice of the initial value x_0 here is inept, the successive approximations $x_n = \varphi(x_{n-1})$ $(n = 1, 2, \ldots)$ can leave the interval (a, b) or even become meaningless. It is therefore useful to have an alternative statement of Theorem 1.

Theorem 2. *Let a function* $\varphi(x)$ *be defined and differentiable on some interval* $[a, b]$, *and let the equation*

$$x = \varphi(x) \tag{17}$$

have a root ξ *located in a smaller interval* $[\alpha, \beta]$, *where* $\alpha = a + \frac{1}{3}(b - a)$ *and* $\beta = b - \frac{1}{3}(b - a)$ (Fig. 33).

Fig. 33



In this case if (*a*) $|\varphi'(x)| \leqslant q < 1$ *for* $a < x < b$ *and* (*b*) *the initial approximation* $x_0 \in [\alpha, \beta]$, *then:*

(1) *all successive approximations lie in the interval* (a, b):

$$x_n = \varphi(x_{n-1}) \in (a, b) \qquad (n = 1, 2, \ldots)$$

(2) *the process of successive approximations is convergent; that is, there exists*

$$\lim_{n \to \infty} x_n = \xi$$

and ξ *is the only root on the interval* $[a, b]$ *of equation* (17), *and*
(3) *estimate* (15) *is valid.*

Proof. (1) Indeed, suppose

$$x_0 \in [\alpha, \beta]$$

Then the equation

$$x_1 = \varphi(x_0)$$

is obviously meaningful. Utilizing the equation

$$\xi = \varphi(\xi)$$

we get, on the basis of the mean-value theorem,

$$|x_1 - \xi| = |\varphi(x_0) - \varphi(\xi)| = |x_0 - \xi|\,|\varphi'(\bar{x}_0)| \leqslant q(\beta - \alpha) < \frac{b-a}{3}$$

whence

$$x_1 \in (a, b)$$

Generally, if $x_{n-1} \in (a, b)$ $(n = 1, 2, \ldots)$ and $|x_{n-1} - \xi| < \frac{b-a}{3}$, then

$$x_n = \varphi(x_{n-1})$$

is meaningful and

$$|x_n - \xi| = |\varphi(x_{n-1}) - \varphi(\xi)| = |x_{n-1} - \xi|\,|\varphi'(\bar{x}_{n-1})| \leqslant$$
$$\leqslant q\,|x_{n-1} - \xi| < \frac{b-a}{3}$$

Consequently, $x_n \in (a, b)$ where $n = 1, 2, 3, \ldots$.

The proofs of assertions (2) and (3) are completely analogous to the proof of Theorem 1.

Note. Suppose that in a certain neighbourhood (a, b) of the root ξ of equation (17), the derivative $\varphi'(x)$ **preserves sign** and the inequality

$$|\varphi'(x)| \leqslant q < 1$$

is valid.

Then if the derivative $\varphi'(x)$ *is positive, the successive approximations*

$$x_n = \varphi(x_{n-1}) \quad (n = 1, 2, \ldots), \quad x_0 \in (a, b)$$

converge to the root ξ *monotonically.*

However, if the derivative $\varphi'(x)$ *is negative, then the successive approximations oscillate about the root* ξ.

(1) Indeed, let $0 \leqslant \varphi'(x) \leqslant q < 1$ and, say,

$$x_0 < \xi$$

Then

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\,\varphi'(\xi_1) < 0$$

where $\xi_1 \in (x_0, \xi)$, and

$$|x_1 - \xi| \leqslant q\,|x_0 - \xi| < |x_0 - \xi|$$

Consequently

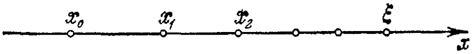
$$x_0 < x_1 < \xi$$

Using the method of mathematical induction, we obtain

$$x_0 < x_1 < x_2 < \ldots < \xi$$

(Fig. 34a).

Fig. 34a



A similar result is obtained when $x_0 > \xi$.

Thus, if the derivative $\varphi'(x)$ is positive, it suffices only to choose the initial approximation x_0 belonging to the neighbourhood (a, b) of the root ξ that interests us; all the remaining approximations $x_n (n = 1, 2, \ldots)$ will automatically lie in this neighbourhood and will monotonically approach the root ξ as n increases.

(2) Let $-1 < -q \leqslant \varphi'(x) \leqslant 0$ and, say, $x_0 < \xi$; $x_1 = \varphi(x_0) \in (a, b)$.

We have

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\,\varphi'(\xi_1) > 0$$

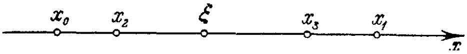
that is, $x_1 > \xi$ and $|x_1 - \xi| < |x_0 - \xi|$.

Repeating these arguments for the approximations $x_1, x_2, \ldots,$ we get

$$x_0 < x_2 < \ldots < \xi < \ldots < x_3 < x_1$$

Thus, the successive approximations oscillate about the root ξ (Fig. 34b).

Fig. 34b



Thus, in the case of a negative derivative $\varphi'(x)$, if two approximations x_0 and x_1 belong to the neighbourhood (a, b) of the root ξ, all the other approximations $x_n (n = 2, 3, \ldots)$ also belong to this neighbourhood; the sequence $\{x_n\}$ "strangles" the root ξ.

Note that, obviously,

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|$$

that is, in this case the stabilized digits of the approximation x_n definitely belong to the exact root ξ.

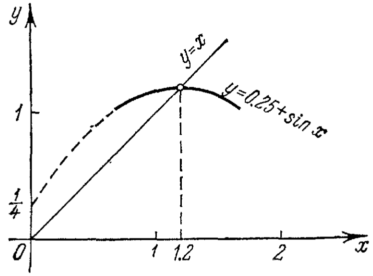
Example 1. Find the real roots of the equation $x - \sin x = 0.25$ to three significant digits.

Solution. Write the equation as

$$x = \sin x + 0.25$$

We establish graphically that the equation has one real root ξ approximately equal to $x_0 = 1.2$ in the interval $[1.1, 1.3]$ (Fig. 35).

Fig. 35



Using the notation of Theorem 2, we put

$$\alpha = 1.1 \quad \text{and} \quad \beta = 1.3$$

whence

$$a = \alpha - (\beta - \alpha) = 0.9 \approx \text{arc}\, 52^\circ$$

and

$$b = \beta + (\beta - \alpha) = 1.5 \approx \text{arc}\, 86^\circ$$

Since

$$\varphi(x) = \sin x + 0.25$$

and

$$\varphi'(x) = \cos x$$

we have, for $0.9 < x < 1.5$,

$$|\varphi'(x)| \leqslant \cos 52^\circ \approx 0.62 = q$$

If we choose $x_0 \in (1.1, 1.3)$, then all the conditions of Theorem 2 will be obeyed and, hence, we will be assured that the successive approximations

$$x_n = \sin x_{n-1} + 0.25 \qquad (n = 1, 2, \ldots)$$

(1) lie in the interval $(0.9, 1.5)$ and (2) $x_n \to \xi$ as $n \to \infty$.

Choosing $x_0 = 1.2$ and specifying, according to the hypothesis of the problem, the limiting absolute error

$$\varepsilon = \frac{1}{2} \cdot 10^{-2}$$

we construct the successive approximations x_n $(n = 1, 2, \ldots)$ until two adjacent approximations x_{n-1} and x_n coincide to within the limits of error equal to

$$\frac{1-q}{q}\varepsilon = 0.51 \cdot \frac{1}{2} \cdot 10^{-2} \approx 0.0025$$

We have

$$\begin{aligned}
x_1 &= \sin 1.2 + 0.25 = 0.932 + 0.25 = 1.182,\\
x_2 &= \sin 1.182 + 0.25 = 0.925 + 0.25 = 1.175,\\
x_3 &= \sin 1.175 + 0.25 = 0.923 + 0.25 = 1.173,\\
x_4 &= \sin 1.173 + 0.25 = 0.922 + 0.25 = 1.172,\\
x_5 &= \sin 1.172 + 0.25 = 0.922 + 0.25 = 1.172
\end{aligned}$$

The fourth and fifth approximations coincide to within four significant digits. Therefore [see (16″)]

$$|x_5 - \xi| \leqslant \frac{0.62 \cdot 0.001}{1-0.62} = 0.0016$$

Since the limiting absolute error of the approximate root x_5 (including the rounding error) does not exceed

$$E = 0.0016 + 0.002 < \frac{1}{2} \cdot 10^{-2}$$

we can take it that

$$\xi = 1.17 \pm 0.005$$

Note. The given equation

$$f(x) = 0 \tag{18}$$

may be written as

$$x = \varphi(x) \tag{18′}$$

choosing the function $\varphi(x)$ in different ways.

The notation of (18′) is by no means immaterial; in some cases $|\varphi'(x)|$ will prove to be small in the neighbourhood of the desired root ξ, in others it will be large. For the method of iteration, the most advantageous representation of (18′) is that in which the inequality

$$|\varphi'(x)| \leqslant q < 1 \tag{19}$$

is valid, and the smaller the number q, the faster, generally speaking, will the successive approximations converge to the root ξ.

We give here one rather general technique for reducing equation (18) to the form (18′), for which the validity of inequality (19) is ensured. Let the desired root ξ of the equation lie in the interval $[a, b]$, and

$$0 < m_1 \leqslant f'(x) \leqslant M_1 \tag{20}$$

for $a \leqslant x \leqslant b$.¹⁾ In particular, we can take for m_1 the smallest

---

¹⁾ If the derivative $f'(x)$ is negative, then we consider the equation $-f(x) = 0$ instead of $f(x) = 0$.

value of the derivative $f'(x)$ on the interval $[a, b]$, which value must be positive, and for M_1 the greatest value of $f'(x)$ on the interval $[a, b]$. Replace (18) by the equivalent equation

$$x = x - \lambda f(x) \quad (\lambda > 0)$$

We can set $\varphi(x) = x - \lambda f(x)$.

We choose the parameter λ in such a way that in the given neighbourhood $[a, b]$ of the root ξ the inequality

$$0 \leqslant \varphi'(x) = 1 - \lambda f'(x) \leqslant q < 1 \tag{21}$$

is valid, whence, on the basis of expression (20), we get

$$0 \leqslant 1 - \lambda M_1 \leqslant 1 - \lambda m_1 \leqslant q$$

Consequently, we can choose

$$\lambda = \frac{1}{M_1}$$

and

$$q = 1 - \frac{m_1}{M_1} < 1$$

Inequality (21) is thus valid.

Example 2. Find the largest positive root ξ of the equation

$$x^3 + x = 1000 \tag{22}$$

to within 10^{-4}.

Solution. A rough guess gives us the approximate value of the root $x_0 = 10$; it is obvious that $\xi < x_0$.

Equation (22) may be written in the form

$$x = 1000 - x^3 \tag{22'}$$

or

$$x = \frac{1000}{x^2} - \frac{1}{x} \tag{22''}$$

or

$$x = \sqrt[3]{1000 - x} \tag{22'''}$$

and so forth. The most suitable version is (22‴) because by taking the interval (9,10) for the main interval and setting

$$\varphi(x) = \sqrt[3]{1000 - x}$$

we have

$$\varphi'(x) = \frac{-1}{3\sqrt[3]{(1000 - x)^2}}$$

whence

$$|\varphi'(x)| \leqslant \frac{1}{3\sqrt[3]{990^2}} \approx \frac{1}{300} = q$$

Compute the successive approximations x_n with one extra digit using the formulas

$$y_n = 1000 - x_n,$$
$$x_{n+1} = \sqrt[3]{y_n} \quad (n = 0, 1, 2, \ldots)$$

The values thus found are listed in Table 4.

TABLE 4

VALUES OF THE SUCCESSIVE APPROXIMATIONS x_n AND y_n

n	x_n	y_n
0	10	990
1	9.96655	990.03345
2	9.96666	990.03334
3	9.96667	

Since $1 - q \approx 1$, we can put $\xi = 9.9667$ to within an accuracy of 10^{-4}.

The method of iteration can also be used to compute the roots of equations given in the form of power series.

Example 3. Find the real root of the equation [2]

$$x - \frac{x^3}{3} + \frac{x^5}{10} - \frac{x^7}{42} + \frac{x^9}{216} - \frac{x^{11}}{1320} + \ldots$$
$$\ldots + (-1)^{n-1} \frac{x^{2n-1}}{(n-1)!\,(2n-1)} + \ldots = 0.4431135$$

Solution. We have $x = \varphi(x)$, where

$$\varphi(x) = 0.4431135 + \frac{x^3}{3} - \frac{x^5}{10} + \frac{x^7}{42} - \frac{x^9}{216} + \frac{x^{11}}{1320} - \ldots$$

Neglecting all powers of x above the first, we determine an approximate value x_0 of the root to be 0.44. Then

$$x_1 = \varphi(0.44) \approx 0.47,$$
$$x_2 = \varphi(0.47) \approx 0.476,$$
$$x_3 = \varphi(0.476) \approx 0.4767,$$
$$x_4 = \varphi(0.4767) \approx 0.47689,$$

$$x_5 = \varphi(0.47689) \approx 0.476927,$$
$$x_6 = \varphi(0.476927) \approx 0.476934,$$
$$x_7 = \varphi(0.476934) \approx 0.476936$$

Hence, $\xi = 0.47693$.

We give another technique for accelerating the convergence of the process of iteration which may be useful in certain cases [7]. Suppose we have an equation

$$x = \varphi(x)$$

such that the inequality

$$|\varphi'(x)| \geqslant k > 1$$

holds within the neighbourhood of the desired root ξ. Then the process of iteration will diverge for this equation. But if the given equation is replaced by the equivalent equation

$$x = \psi(x)$$

where $\psi(x) = \varphi^{-1}(x)$ is the inverse function, we get an equation for which the process of iteration converges since

$$|\psi'(x)| = \left|\frac{1}{\varphi'(\psi(x))}\right| \leqslant \frac{1}{k} = q < 1$$

Example 4. The equation

$$f(x) \equiv x^3 - x - 1 = 0 \tag{23}$$

has a root $\xi \in (1, 2)$ since $f(1) = -1 < 0$ and $f(2) = 5 > 0$.

Equation (23) may be written as

$$x = x^3 - 1 \tag{24}$$

Here

$$\varphi(x) = x^3 - 1 \quad \text{and} \quad \varphi'(x) = 3x^2$$

and so

$$\varphi'(x) \geqslant 3 \quad \text{for} \quad 1 \leqslant x \leqslant 2$$

and, hence, the conditions of convergence of the process of iteration are not fulfilled.

If we write (23) as

$$x = \sqrt[3]{x+1} \tag{25}$$

we will have

$$\psi(x) = \sqrt[3]{x+1} \quad \text{and} \quad \psi'(x) = \frac{1}{3\sqrt[3]{(x+1)^2}}$$

Whence $0 < \psi'(x) < \frac{1}{3\sqrt[3]{4}} < \frac{1}{4}$ for $1 \leqslant x \leqslant 2$ and, hence, for equation (25) the process of iteration will converge rapidly.

4.9 THE METHOD OF ITERATION FOR A SYSTEM OF TWO EQUATIONS

Let there be given two equations in two unknowns:

$$\left.\begin{aligned} F_1(x, y) = 0, \\ F_2(x, y) = 0 \end{aligned}\right\} \tag{1}$$

whose real roots it is required to find to within a specified degree of accuracy.

We assume that the system (1) allows only for isolated roots. We can establish the number of these roots and give their crudely approximate values by constructing the curves $F_1(x, y) = 0$, $F_2(x, y) = 0$ and determining the coordinates of their points of intersection.

Let $x = x_0$, $y = y_0$ be the approximate values of the roots of system (1) obtained graphically or in some other way (say, by a rough guess).

We offer an iteration process which, under certain circumstances, permits improving the given approximate values of the roots. To do this, represent (1) as

$$\left.\begin{aligned} x = \varphi_1(x, y), \\ y = \varphi_2(x, y) \end{aligned}\right\} \tag{2}$$

and construct the successive approximations according to the following formulas:

$$\begin{array}{ll} x_1 = \varphi_1(x_0, y_0), & y_1 = \varphi_2(x_0, y_0), \\ x_2 = \varphi_1(x_1, y_1), & y_2 = \varphi_2(x_1, y_1), \\ \dots\dots\dots\dots & \dots\dots\dots\dots \\ x_{n+1} = \varphi_1(x_n, y_n), & y_{n+1} = \varphi_2(x_n, y_n), \\ \dots\dots\dots\dots & \dots\dots\dots\dots \end{array} \tag{3}$$

If the iteration process (3) converges, that is, if there exist the limits

$$\xi = \lim_{n\to\infty} x_n \quad \text{and} \quad \eta = \lim_{n\to\infty} y_n$$

then, assuming the functions $\varphi_1(x, y)$ and $\varphi_2(x, y)$ to be continuous and passing to the limit in the general-form equation (3), we get

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \varphi_1(x_n, y_n),$$

$$\lim_{n\to\infty} y_{n+1} = \lim_{n\to\infty} \varphi_2(x_n, y_n)$$

Whence

$$\xi=\varphi_1(\xi,\ \eta),\quad \eta=\varphi_2(\xi,\ \eta)$$

that is the limiting values ξ and η are roots of system (2) and, hence, of system (1) as well. Therefore, taking a sufficiently large number of iterations (3), we get the numbers x_n and y_n which will differ from the exact roots $x=\xi$ and $y=\eta$ of (1) by an arbitrarily small value. The problem is thus solved. If the iteration process (3) diverges, it cannot be used.

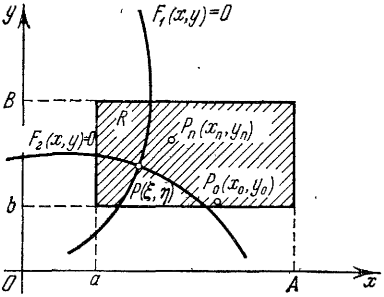


Fig. 36

Theorem. *In a closed neighbourhood* R $\{a \leqslant x \leqslant A;\ b \leqslant y \leqslant B\}$ (Fig. 36) *let there be one and only one pair of roots* $x=\xi$ *and* $y=\eta$ *of system* (2). *If*: (1) *the functions* $\varphi_1(x,\ y)$ *and* $\varphi_2(x,\ y)$ *are defined and continuously differentiable in* R; (2) *the initial approximations* x_0, y_0 *and all succeeding approximations* x_n, y_n $(n=1,\ 2,\ \ldots)$ *belong to* R; (3) *the following inequalities are valid in* R:

$$\left|\frac{\partial\varphi_1}{\partial x}\right|+\left|\frac{\partial\varphi_2}{\partial x}\right|\leqslant q_1<1,$$
$$\left|\frac{\partial\varphi_1}{\partial y}\right|+\left|\frac{\partial\varphi_2}{\partial y}\right|\leqslant q_2<1$$

then the process of successive approximations (3) *converges to the roots* $x=\xi$ *and* $y=\eta$ *of system* (2), *that is,*

$$\lim_{n\to\infty} x_n=\xi \quad \text{and} \quad \lim_{n\to\infty} y_n=\eta$$

Note. The theorem holds true if Condition (3) is replaced by Condition (3′):

$$\left|\frac{\partial\varphi_1}{\partial x}\right|+\left|\frac{\partial\varphi_1}{\partial y}\right|\leqslant q_1<1,$$
$$\left|\frac{\partial\varphi_2}{\partial x}\right|+\left|\frac{\partial\varphi_2}{\partial y}\right|\leqslant q_2<1$$

A rough proof of this theorem is given in [2]. A more general theorem is given in Secs. 13.10 and 13.11.

Example. For the system [2]

$$\left.\begin{aligned} f_1(x,\ y)&\equiv 2x^2-xy-5x+1=0,\\ f_2(x,\ y)&\equiv x+3\log_{10}x-y^2=0\end{aligned}\right\}$$

find the positive roots to four significant digits.

Solution. Plot the graphs of the functions $f_1(x, y)=0$ and $f_2(x, y)=0$ (Fig. 37). The approximate values of the roots that interest us are

$$x_0=3.5, \quad y_0=2.2$$

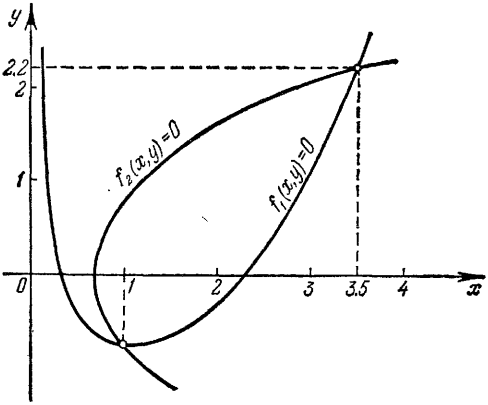


Fig. 37

To apply the method of iteration, write the system as

$$x=\sqrt{\frac{x(y+5)-1}{2}} \equiv \varphi_1(x, y),$$
$$y=\sqrt{x+3\log_{10} x} \equiv \varphi_2(x, y)$$

We find the partial derivatives:

$$\frac{\partial \varphi_1}{\partial x}=\frac{y+5}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial \varphi_2}{\partial x}=\frac{1+\frac{3M}{x}}{2\sqrt{x+3\log_{10} x}}$$

where $M=0.43429$,

$$\frac{\partial \varphi_1}{\partial y}=\frac{x}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial \varphi_2}{\partial y}=0$$

Restricting ourselves to the neighbourhood $R\{|x-3.5| \leqslant 0.1, |y-2.2| \leqslant 0.1\}$ we have

$$\left|\frac{\partial \varphi_1}{\partial x}\right| \leqslant \frac{2.3+5}{4\sqrt{\frac{3.4(2.1+5)-1}{2}}}<0.54,$$

$$\left|\frac{\partial \varphi_1}{\partial y}\right| \leqslant \frac{3.6}{4\sqrt{\frac{3.4(2.1+5)-1}{2}}}<0.27,$$

$$\left|\frac{\partial\varphi_2}{\partial x}\right| \leqslant \frac{1+\frac{3\cdot 0.43}{3.4}}{2\sqrt{3.4+3\log_{10}3.4}} < 0.42,$$

$$\left|\frac{\partial\varphi_2}{\partial y}\right| = 0$$

whence

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial x}\right| < 0.54+0.42=0.96<1, \qquad (4)$$

$$\left|\frac{\partial\varphi_1}{\partial y}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| < 0.27+0=0.27<1 \qquad (5)$$

Thus, if the successive approximations (x_n, y_n) do not leave the region R (this is easy to see as the computations progress), the iteration process will converge.

The relative proximity of the sum (4) to unity permits assuming that the iteration process in this case will converge comparatively slowly. Begin computing the successive approximations by the formulas

$$x_{n+1} = \sqrt{\frac{x_n(y_n+5)-1}{2}},$$

$$y_{n+1} = \sqrt{x_n + 3\log_{10} x_n} \qquad (n=0, 1, 2, \ldots)$$

The corresponding values of the successive approximations are given in Table 5.

TABLE 5

VALUES OF THE SUCCESSIVE APPROXIMATIONS x_n AND y_n

n	x_n	y_n
0	3.5	2.2
1	3.479	2.259
2	3.481	2.260
3	3.484	2.261
4	3.486	2.261
5	3.487	2.262
6	3.487	2.262

We can thus take $\xi = 3.487$ and $\eta = 2.262$.

Note. In place of the above-considered process of successive approximations (3), it is sometimes more convenient to use *Seidel's process:*

$$x_{n+1} = \varphi_1(x_n, y_n),$$

$$y_{n+1} = \varphi_2(x_{n+1}, y_n) \qquad (n=0, 1, 2, \ldots)$$

The method of iteration for general systems is considered in Secs. 13.8 to 13.11.

4.10 NEWTON'S METHOD FOR A SYSTEM OF TWO EQUATIONS

Let x_n, y_n be the approximate roots of the system of equations

$$F(x, y)=0, \quad G(x, y)=0 \tag{1}$$

where F and G are continuously differentiable functions. Putting

$$x=x_n+h_n, \quad y=y_n+k_n$$

we get

$$\left.\begin{aligned} F(x_n+h_n, \ y_n+k_n)=0, \\ G(x_n+h_n, \ y_n+k_n)=0 \end{aligned}\right\} \tag{2}$$

Whence, using Taylor's formula and confining ourselves to linear terms in h_n and k_n, we get

$$\left.\begin{aligned} F(x_n, y_n)+h_nF'_x(x_n, y_n)+k_nF'_y(x_n, y_n)=0, \\ G(x_n, y_n)+h_nG'_x(x_n, y_n)+k_nG'_y(x_n, y_n)=0 \end{aligned}\right\} \tag{3}$$

If the Jacobian

$$J(x_n, y_n)=\begin{vmatrix} F'_x(x_n, y_n) & F'_y(x_n, y_n) \\ G'_x(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix} \neq 0$$

then from system (3) we have

$$h_n=-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{4}$$

$$k_n=-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix} \tag{5}$$

We can thus put

$$x_{n+1}=x_n-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{6}$$

$$y_{n+1}=y_n-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix} \tag{6'}$$

$(n=0, 1, 2, \ldots)$

The initial approximations x_0, y_0 are determined very roughly.

Example. Find the real roots of the system

$$\begin{aligned} F(x, y) &\equiv 2x^3-y^2-1=0, \\ G(x, y) &\equiv xy^3-y-4=0 \end{aligned} \tag{1}$$

Solution. Graphically we obtain crude approximations to the values of the roots:

$$x_0 = 1.2, \quad y_0 = 1.7$$

Substituting them into (1) we get

$$F(1.2,\ 1.7) = -0.434,$$
$$G(1.2,\ 1.7) = 0.1956$$

Compute the Jacobian

$$J(x,\ y) = \begin{vmatrix} 6x^2 & -2y \\ y^3 & 3xy^2 - 1 \end{vmatrix}$$

whence

$$J = \begin{vmatrix} 8.64 & -3.40 \\ 4.91 & 9.40 \end{vmatrix} = 97.910$$

We compute h_0 from formula (4):

$$h_0 = -\frac{1}{97.910} \begin{vmatrix} -0.434 & -3.40 \\ 0.1956 & 9.40 \end{vmatrix} = \frac{3.389}{97.910} = 0.0349$$

and from this, using (6), we obtain

$$x_1 = 1.2 + 0.0349 = 1.2349$$

Compute k_0 from formula (5):

$$k_0 = -\frac{1}{97.910} \begin{vmatrix} 8.64 & -0.434 \\ 4.91 & 0.1956 \end{vmatrix} = -0.0390$$

whence, by (6), we get

$$y_1 = 1.7 - 0.0390 = 1.6610$$

Repeating this process with the roots obtained, we find $x_2 = 1.2343$, $y_2 = 1.6615$ and so on.

Newton's method for general systems is considered in Secs. 13.1 to 13.7.

4.11 NEWTON'S METHOD FOR THE CASE OF COMPLEX ROOTS

In practical situations (for instance when solving linear differential equations) it may be necessary to improve the complex roots of a given equation

$$f(z) = 0 \tag{1}$$

A method similar to Newton's may sometimes be used for this purpose.

Suppose that $f(z)$ $(z=x+iy,\ i^2=-1)$ is an analytic function in some convex[1] neighbourhood U of its simple isolated zero

$$\zeta=\xi+i\eta \qquad (f(\zeta)=0,\ f'(\zeta)\neq 0)$$

which, generally speaking, is complex. Let z_n be an approximate value of the root, which value lies in the neighbourhood U, and let

$$z_{n+1}=z_n+\Delta z_n$$

be an improved value of the root. Using a Taylor series expansion at the point z_n, assuming $f(z_{n+1})\approx 0$ to within Δz_n^2, we have

$$f(z_{n+1})\approx f(z_n)+\Delta z_n f'(z_n)=0$$

whence

$$\Delta z_n=-\frac{f(z_n)}{f'(z_n)} \tag{2}$$

Thus, starting with a value z_0, we can, step by step, obtain subsequent approximations of the root using the formula

$$z_{n+1}=z_n-\frac{f(z_n)}{f'(z_n)} \qquad (n=0,\ 1,\ 2,\ \ldots) \tag{3}$$

If $z_n\in U$ $(n=1, 2, \ldots)$ and the sequence $\{z_n\}$ converges, then the limit

$$\zeta=\lim_{n\to\infty} z_n$$

is a root of equation (1). Indeed, passing to the limit in (3) as $n\to\infty$, we have

$$\lim_{n\to\infty} z_{n+1}=\lim_{n\to\infty} z_n-\frac{\lim\limits_{n\to\infty} f(z_n)}{\lim\limits_{n\to\infty} f'(z_n)}$$

or

$$\zeta=\zeta-\frac{f(\zeta)}{f'(\zeta)}$$

Consequently

$$f(\zeta)=0$$

To estimate the error in the approximate value z_n, assume that

$$|f'(z)|\geqslant m_1>0 \quad \text{for} \quad z\in U$$

Then for the given function

$$w=f(z)$$

---

[1] That is, any two points in U are endpoints of a segment which is also in U.

there exists, in a sufficiently small R-neighbourhood of the root ζ, a single-valued inverse function

$$z = f^{-1}(w)$$

defined in a neighbourhood $|w| < \rho$, the derivative of which is, as we know,

$$\frac{dz}{dw} = \frac{1}{f'(z)} \tag{4}$$

Assuming that $|f(z_n)| < \rho$ we have

$$z_n - \zeta = f^{-1}(f(z_n)) - f^{-1}(f(\zeta)) = \int_{f(\zeta)}^{f(z_n)} \frac{d}{dt}[f^{-1}(t)]\,dt = \int_0^{f(z_n)} \frac{dt}{f'(f^{-1}(t))} \tag{5}$$

where t is a point ranging over the rectilinear segment between the points $f(\zeta) = 0$ and $f(z_n)$ (Fig. 38). Since $|t| < \rho$, it follows that $|f^{-1}(t)| < R$ and, hence,

$$|f'(f^{-1}(t))| \geqslant m_1$$

From this, on the basis of (5), we get

$$|z_n - \zeta| \leqslant \int_0^{f(z_n)} \frac{|dt|}{|f'(f^{-1}(t))|} \leqslant \frac{|f(z_n)|}{m_1} \tag{6}$$

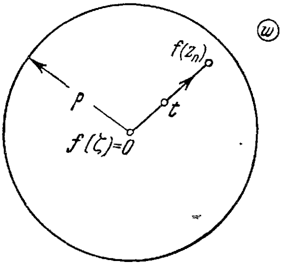


Fig. 38

We give without proof the sufficient conditions for the existence of a root of equation (1) which follow from the Ostrowski theorem.

Theorem. *If a function* f(z) *is analytic in a closed* R*-neighbourhood of a point* z_0, *and the following inequalities hold:*

(1) $\left|\frac{1}{f'(z_0)}\right| \leqslant A_0$,

(2) $\left|\frac{f(z_0)}{f'(z_0)}\right| \leqslant B_0 \leqslant \frac{R}{2}$,

(3) $|f''(z)| \leqslant C$ for $|z - z_0| < R$,

(4) $2A_0B_0C = \mu_0 \leqslant 1$

then equation (1) *has a unique root* ζ *in the domain* $|z - z_0| \leqslant R$ *and Newton's process* (3) *defined by the initial approximation* z_0 *converges to this root, that is,*

$$\zeta = \lim_{n \to \infty} z_n.$$

The rapidity of convergence of the process is characterized by the estimate

$$|\zeta - z_n| \leqslant B_0 \left(\frac{1}{2}\right)^{n-1} \mu_0^{2^n - 1} \tag{7}$$

Example. Find approximately the smallest, in modulus, roots of the equation

$$f(z) \equiv e^z - 0.2z + 1 = 0 \tag{8}$$

Solution. Here

$$f'(z) = e^z - 0.2$$

Since $f'(z) = 0$ for $\tilde{z} = \ln 0.2 \approx -1.79$ and

$$f(-\infty) = +\infty, \quad f(\tilde{z}) > 0, \quad f(+\infty) = +\infty$$

it follows that equation (8) does not have any real roots.

For the initial approximation of the desired root ζ we take the smallest, in modulus, root z_0 of the equation

$$e^z + 1 = 0$$

whence we can put

$$z_0 = \pi i$$

The succeeding approximations z_n $(n = 1, 2, 3, \ldots)$ of the root ζ are successively found from formula (3):

$$z_1 = z_0 - \frac{f(z_0)}{f'(z_0)} = \pi i - \frac{0.2\pi i}{1.2} = \frac{5}{6}\pi i = 2.618i,$$

$$z_2 = z_1 - \frac{f(z_1)}{f'(z_1)} = \frac{5\pi i}{6} - \frac{0.132 - 0.024i}{-1.868 + 0.5i} = 0.069 + 2.624i, \text{ etc.}$$

The results of the computations to within 0.001 are given in Table 6.

TABLE 6

IMPROVING COMPLEX ROOTS BY NEWTON'S METHOD

n	z_n	e^{z_n}	$f(z_n)$	$f'(z_n)$	$\Delta z_n = -\frac{f(z_n)}{f'(z_n)}$
0	$3.142i$	-1	$-0.628i$	-1.2	$-0.524i$
1	$2.618i$	$-0.868 + 0.5i$	$0.132 - 0.024i$	$-1.068 + 0.5i$	$0.153 + 0.040i$
2	$0.153 + 2.658i$	$-1.030 + 0.541i$	$-0.061 + 0.009i$	$-1.230 + 0.541i$	$-0.044 - 0.012i$
3	$0.109 + 2.646i$	$-0.978 + 0.535i$	$0 + 0.006i$	$-1.178 + 0.535i$	$-0.002 + 0.004i$
4	$0.107 + 2.650i$	$-0.981 + 0.525i$	$-0.002 - 0.005i$	$-1.181 + 0.525i$	$-0.000 - 0.004i$
5	$0.107 + 2.646i$	$-0.977 + 0.534i$	$+0.002 + 0.004i$	$-1.177 + 0.534i$	

To compute e^z for $z = x + iy$ we used the familiar formula

$$e^z = e^x(\cos y + i \sin y)$$

Assuming

$$\zeta \approx z_5 = 0.107 + 2.646i$$

we have

$$f(z_5) = 0.002 + 0.004i$$

Approximately taking it that

$$m_1 = |f'(z_5)| \approx 1.3$$

we obtain the error on the basis of (6):

$$|\zeta - z_5| \approx \frac{|f(z_5)|}{m_1} = \frac{0.001 \cdot \sqrt{20}}{1.3} \approx 0.004$$

Since the left member of (8) takes on real values for real z, this equation also has the conjugate root

$$\bar{\zeta} \approx 0.107 - 2.646i$$

which is equal, in modulus, to the root ζ. Indeed, we have

$$f(\bar{\zeta}) = \overline{f(\zeta)} = 0$$

Note. An alternative method of solving (1) is to reduce it to a system of two real equations. Setting

$$z = x + iy$$

in (1) and isolating the real and imaginary parts of the function $f(z)$, we get

$$f(z) \equiv u(x, y) + iv(x, y) = 0$$

where u and v are real functions. From this we find that (1) is equivalent to the system

$$\left.\begin{aligned} u(x, y) &= 0, \\ v(x, y) &= 0 \end{aligned}\right\} \tag{9}$$

Improvement of the roots of a system of type (9) is considered in Secs. 4.9 and 4.10. Note also that this new method is suitable for the case when the function $f(z)$ is nonanalytic.

REFERENCES FOR CHAPTER 4

[1] *Ya. S. Bezikovich, Approximate Computations*, 1949, Chapter VI (in Russian).
[2] *J. B. Scarborough, Numerical Mathematical Analysis*, 1955, Chapters IX, X.
[3] *E. T. Whittaker and G. Robinson, The Calculus of Observations*, 1944, Chapter VI.
[4] *G. M. Fikhtengolts, Course of Differential and Integral Calculus*, 1957, Vol. I, Chapter IV (in Russian).
[5] *G. P. Tolstov, Course of Mathematical Analysis*, 1954, Vol. I, Chapter VII (in Russian).
[6] *A. O. Gelfond, Calculus of Finite Differences*, 1952, Chapter V (in Russian).
[7] *D. A. Ventsel, E. S. Ventsel, Elements of the Theory of Approximate Computations*, 1949, Chapter 3, Sec. 4 (in Russian).
[8] *L. V. Kantorovich, On Newton's Method*, 1949 (in Russian).

Chapter 5

SPECIAL TECHNIQUES FOR APPROXIMATE SOLUTION OF ALGEBRAIC EQUATIONS

5.1 GENERAL PROPERTIES OF ALGEBRAIC EQUATIONS

Consider the *algebraic equation* of degree $n\ (n \geqslant 1)$

$$P(x) \equiv a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

where the coefficients $a_0, a_1, \ldots, a_n$ are real numbers and

$$a_0 \neq 0$$

The variable x will be considered complex in the general case.

Fundamental theorem of algebra. *An algebraic equation of the* n*th degree,* (1), [*and, hence, also the polynomial* P (x)] *has exactly* n *roots, real or complex, provided that each root is counted according to its multiplicity* [1], [2].

We then say that a root ξ of equation (1) has multiplicity s (that is, ξ is an s-*fold root*) if

$$\begin{gathered} P(\xi) = P'(\xi) = \ldots = P^{(s-1)}(\xi) = 0, \\ P^{(s)}(\xi) \neq 0 \end{gathered} \tag{2}$$

The complex roots of equation (1) have the property of appearing in *complex conjugate pairs.*

Theorem 1. *If the coefficients of the algebraic equation* (1) *are real, then the complex roots of this equation are complex conjugate in pairs, that is if* $\xi = \alpha + i\beta$ (α, β are real) *is a root of* (1), *of multiplicity* s, *then the number* $\bar{\xi} = \alpha - i\beta$ *is also a root of that equation and has the same multiplicity* s.

Note that the moduli of these roots are the same:

$$|\xi| = |\bar{\xi}| = \sqrt{\alpha^2 + \beta^2}$$

Corollary. *An algebraic equation of odd degree with real coefficients has at least one real root.*

It is easy to give a rough estimate of the moduli of the roots of equation (1).

Theorem 2. *Suppose*

$$A=\max\{|a_1|,\ |a_2|,\ \ldots,\ |a_n|\}$$

where a_k *are the coefficients of* (1).

Then the moduli of all the roots x_k $(k=1,\ \ldots,\ n)$ *of* (1) *satisfy the inequality*

$$|x_k|<1+\frac{A}{|a_0|} \qquad (3)$$

That is, in the complex plane $\xi O\eta$ $(x=\xi+i\eta)$ *the roots of this equation are located inside the circle*

$$|x|<1+\frac{A}{|a_0|}=R$$

(Fig. 39).

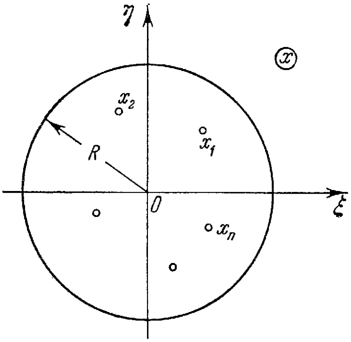


Fig. 39

Proof. Setting $|x|>1$, we have from formula (1)

$$\begin{aligned}|P(x)|&\geqslant|a_0x^n|-(|a_1x^{n-1}|+|a_2x^{n-2}|+\ldots+|a_n|)\geqslant\\&\geqslant|a_0||x|^n-A(|x|^{n-1}+|x|^{n-2}+\ldots+1)=\\&=|a_0||x|^n-A\frac{|x|^n-1}{|x|-1}>\left(|a_0|-\frac{A}{|x|-1}\right)|x|^n\end{aligned}$$

Whence, if

$$|a_0|-\frac{A}{|x|-1}\geqslant 0$$

that is, if

$$|x|\geqslant 1+\frac{A}{|a_0|} \qquad (4)$$

we find that

$$|P(x)|>0$$

Thus the values of x which satisfy (4) are definitely not the roots of equation (1). Hence, all the roots x_k of (1) satisfy the reversed inequality

$$|x_k|<1+\frac{A}{|a_0|}$$

Corollary. *Let* $a_n\neq 0$ *and*

$$B=\max\{|a_0|,\ |a_1|,\ \ldots,\ |a_{n-1}|\}$$

Then all the roots x_k $(k=1,\ 2,\ \ldots,\ n)$ *of equation* (1) *satisfy the inequality*

$$|x_k|>\frac{1}{1+\frac{B}{|a_n|}}=r \qquad (5)$$

that is, the roots of (1) *are located in the annulus*

$$r < |x| < R$$

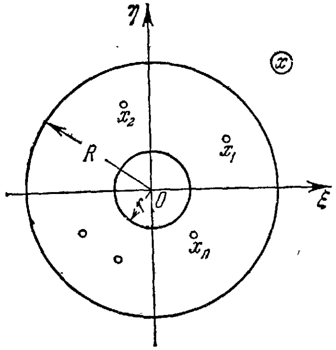


Fig. 40

(Fig. 40).

True enough, for, setting

$$x = \frac{1}{y}$$

we have

$$P(x) = \frac{1}{y^n} Q(y)$$

where

$$Q(y) = a_n y^n + a_{n-1} y^{n-1} + \ldots + a_0$$

The roots $y_k = \frac{1}{x_k}$ $(k = 1, \ldots, n)$ of the polynomial $Q(y)$ satisfy, by virtue of our theorem, the inequality

$$|y_k| = \frac{1}{|x_k|} < 1 + \frac{B}{|a_n|}$$

whence

$$|x_k| > \frac{1}{1 + \frac{B}{|a_n|}} = r \qquad (k = 1, \ldots, n)$$

Note. The numbers r and R are the *upper* and *lower* bounds, respectively, of the positive roots of equation (1).

Similarly, the numbers $-R$ and $-r$ serve as the lower and upper bounds, respectively, of the negative roots of equation (1).

If

$$x_1, x_2, \ldots, x_n$$

are the roots of (1), then the following expansion holds for the left member:

$$P(x) = a_0 (x - x_1)(x - x_2) \ldots (x - x_n) \tag{6}$$

From this, multiplying out the binomials in (6) and equating the coefficients of identical powers of x in the left and right members of (6), we get relations between the roots and the coefficients of the algebraic equation:

$$\left.\begin{array}{l} x_1 + x_2 + \ldots + x_n = -\frac{a_1}{a_0}, \\ x_1 x_2 + x_1 x_3 + \ldots + x_{n-1} x_n = \frac{a_2}{a_0}, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_1 x_2 \ldots x_n = (-1)^n \frac{a_n}{a_0} \end{array}\right\} \tag{7}$$

The left members of (7) are the sums of combinations of the roots of equation (1) taken one at a time, two at a time, etc.

Example 1. The roots x_1, x_2, x_3 of the cubic equation

$$x^3+px^2+qx+r=0$$

satisfy the conditions

$$\left.\begin{array}{l} x_1+x_2+x_3=-p, \\ x_1x_2+x_1x_3+x_2x_3=q, \\ x_1x_2x_3=-r \end{array}\right\}$$

If multiplicities of the roots are taken into account, the expansion (6) assumes the form

$$P(x)=a_0(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2}\dots(x-x_m)^{\alpha_m}$$

where $x_1, x_2, \dots, x_m$ $(m\leqslant n)$ are distinct roots of equation (1) and $\alpha_1, \alpha_2, \dots, \alpha_m$ are their multiplicities, and

$$\alpha_1+\alpha_2+\dots+\alpha_m=n$$

The derivative $P'(x)$ is expressed as follows:

$$P'(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\dots(x-x_m)^{\alpha_m-1}\,Q(x)$$

where $Q(x)$ is a polynomial such that

$$Q(x_k)\neq 0 \quad \text{for} \quad k=1, 2, \dots, m$$

For this reason, polynomial

$$R(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\dots(x-x_m)^{\alpha_m-1}$$

is the greatest common divisor of the polynomial $P(x)$ and its derivative $P'(x)$. It will be recalled that $R(x)$ can be found by Euclid's algorithm [1]. Forming the quotient

$$f(x)=\frac{P(x)}{R(x)}$$

we get the polynomial

$$f(x)=A_0x^m+A_1x^{m-1}+\dots+A_m \tag{8}$$

with real coefficients $A_0=a_0$, $A_1, \dots, A_m$, the roots of which $x_1, x_2, \dots, x_m$ are distinct.

Thus, the solution of an algebraic equation with multiple roots reduces to that of an algebraic equation of lower degree with distinct roots.

The total number of roots $x_1, x_2, \dots, x_N$ of the equation

$$P(x)=0$$

located in the complex plane inside a simple closed contour Γ

(Fig. 41) may be determined on the basis of the *principle of the argument* [4] which consists in the following: *if a polynomial* P (x) *has no roots on a closed contour* Γ, *then the number of roots* N *of the polynomial inside the contour* Γ *is exactly equal to the variation of* Arg P (x) *in a positive traversal of* Γ, *divided by* 2π; that is,

$$N = \frac{1}{2\pi} \Delta_\Gamma \operatorname{Arg} P(x)$$

Each root is counted according to its multiplicity.

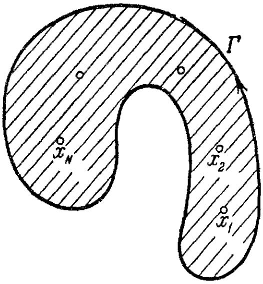


Fig. 41

If the equation of the contour Γ is

$$x = \xi(t) + i\eta(t) \qquad (0 \leqslant t \leqslant T)$$

(t a parameter), then, to determine the number N, one constructs in the xy-plane a curve

$$X = X(t), \quad Y = Y(t) \qquad (0 \leqslant t \leqslant T) \tag{K}$$

where

$$P(x) \doteq P(\xi(t) + i\eta(t)) = X(t) + iY(t)$$

and $X(t)$, $Y(t)$ are real functions; then one counts the number of circuits N that the curve K makes about the origin.

Example 2. Determine the number of roots of the equation

$$P(x) \equiv x^3 - 3x + 1 = 0 \tag{9}$$

contained in the circle $|x| < 2$.

Solution. Putting

$$x = 2(\cos t + i \sin t)$$

we have

$$P(x) = 8(\cos t + i \sin t)^3 - 6(\cos t + i \sin t) + 1 = \\ = (8 \cos 3t - 6 \cos t + 1) + i(8 \sin 3t - 6 \sin t)$$

whence

$$\left.\begin{aligned} X &= 8 \cos 3t - 6 \cos t + 1, \\ Y &= 8 \sin 3t - 6 \sin t \end{aligned}\right\} \tag{K}$$

TABLE 7

t	0	$\pm\frac{\pi}{6}$	$\pm\frac{\pi}{3}$	$\pm\frac{\pi}{2}$	$\pm\frac{2\pi}{3}$	$\pm\frac{5\pi}{6}$	$\pm\pi$
X	3	−4.22	−10	1	15	6.22	−1
Y	0	±5	±5.22	∓14	∓5.22	±5	0

Plotting the curve K (see Table 7), it is easy to see that the curve circles the origin three times (Fig. 42). Therefore, $N=3$ and so equation (9) has three roots inside the circle $|x|<2$.

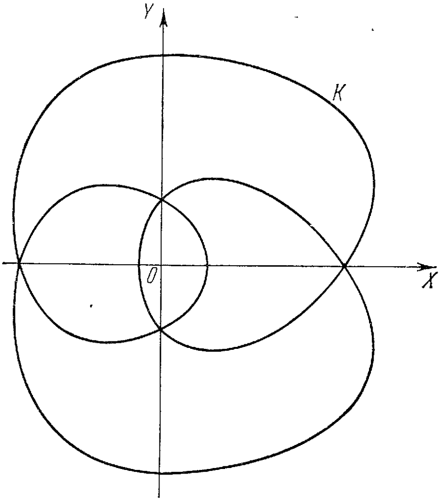


Fig. 42

5.2 THE BOUNDS OF REAL ROOTS OF ALGEBRAIC EGUATIONS

In this section we consider polynomials of the type

$$P(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n \qquad (a_0\neq 0) \tag{1}$$

with real coefficients $a_0, a_1, \ldots, a_n$. Our aim here is to establish the limits, as narrow as possible, for the positive and negative roots $x_1, x_2, \ldots, x_m$ $(1\leqslant m\leqslant n)$ of the equation

$$P(x)=0 \tag{2}$$

The problem of the existence of these roots is not touched on here. It will be noted that one can restrict himself to finding the upper limit R of only the positive roots of equations of type (2). Indeed, along with (2) let us consider the auxiliary algebraic equations

$$P_1(x)\equiv x^nP\left(\frac{1}{x}\right)=0,$$
$$P_2(x)\equiv P(-x)=0,$$
$$P_3(x)\equiv x^nP\left(-\frac{1}{x}\right)=0$$

and let the upper bounds of their positive roots be R_1, R_2 and R_3 respectively. Then the number $\frac{1}{R_1}$ is clearly a lower bound of the positive roots of equation (2), that is, all positive roots x^+ of this equation, if they exist, satisfy the inequality

$$\frac{1}{R_1} \leqslant x^+ \leqslant R$$

Similarly, the numbers $-R_2$ and $-\frac{1}{R_3}$ are, respectively, lower and upper bounds of the negative roots of (2), that is, all negative roots x^- of this equation, if there are any, satisfy the inequality

$$-R_2 \leqslant x^- \leqslant -\frac{1}{R_3}$$

We now give some simple techniques (not all of which are provided with proof) for finding the upper bound R of positive roots of equation (2).

Lagrange's theorem. *Suppose* $a_0 > 0$ *and* a_k $(k \geqslant 1)$ *is the first of the negative coefficients*[1] *of the polynomial* $P(x)$. *Then for the upper bound of the positive roots of* (2) *we can take the number*

$$R = 1 + \sqrt[k]{\frac{B}{a_0}} \tag{3}$$

where B *is the largest absolute value of the negative coefficients of the polynomial* $P(x)$.

Proof. Set $x > 1$. If in $P(x)$ each of the nonnegative coefficients $a_1, \ldots, a_{k-1}$ is replaced by zero, and each of the remaining coefficients $a_k, a_{k+1}, \ldots, a_n$ is replaced by a negative number $-B$, then polynomial (1) can only diminish its value and we will have the inequality

$$P(x) \geqslant a_0x^n - B(x^{n-k} + x^{n-k-1} + \ldots + 1) = a_0x^n - B\frac{x^{n-k+1} - 1}{x-1}$$

Whence for $x > 1$ we have

$$P(x) > a_0x^n - \frac{B}{x-1}x^{n-k+1} = \frac{x^{n-k+1}}{x-1}[a_0x^{k-1}(x-1) - B] >$$

$$> \frac{x^{n-k+1}}{x-1}[a_0(x-1)^k - B]$$

Consequently for

$$x \geqslant 1 + \sqrt[k]{\frac{B}{a_0}} = R$$

we have

$$P(x) > 0$$

---

[1] If there is no such coefficient, i.e. all coefficients of $P(x)$ are nonnegative, then $P(x)$ has no positive roots.

Thus all the positive roots x^+ of equation (2) satisfy the inequality

$$x^+ < R$$

5.3 THE METHOD OF ALTERNATING SUMS

The idea of Lagrange's method may be generalized in the following manner: let a polynomial $P(x)$ be arranged in descending powers of the variable x, its leading coefficient being $a_0 > 0$. Represent $P(x)$ in the form of an alternating sum:

$$P(x) = Q_1(x) - Q_2(x) + Q_3(x) - Q_4(x) + \dots + Q_{2m-1}(x) - Q_{2m}(x)$$

where $Q_1(x)$ is the sum of the successive terms of the polynomial $P(x)$ with positive coefficients beginning with a_0x^n and $-Q_2(x)$ is the sum of the successive terms of the polynomial $P(x)$ with negative coefficients, which terms adjoin the terms of the first sum, and so on; the last summand, $-Q_{2m}(x)$, either consists of terms with negative coefficients or is identically zero.

Denote by c_j $(j = 1, 2, \dots, m)$ positive numbers such that

$$Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0 \tag{1}$$

$(j = 1, 2, \dots, m)$. Then for the upper bound of the positive roots of equation (2) of Sec. 5.2 we can take the number

$$R = \max(c_1, c_2, \dots, c_m) \tag{2}$$

True enough, set

$$Q_{2j-1}(x) - Q_{2j}(x) = b_1^{(j)}x^{n_j} + b_2^{(j)}x^{n_j-1} + \dots + b_p^{(j)}x^{n_j-p+1} -$$
$$- b_{p+1}^{(j)}x^{n_j-p} - b_{p+2}^{(j)}x^{n_j-p-1} - \dots - b_{p+q}^{(j)}x^{n_j-p-q+1}$$

where

$$b_s^{(j)} \geqslant 0 \qquad (s = 1, 2, \dots, p+q)$$

and

$$b_1^{(j)} > 0 \qquad (j = 1, 2, \dots, m)$$

Assuming $x > 0$, we have

$$Q_{2j-1}(x) - Q_{2j}(x) = x^{n_j-p+1}\left[\left(b_1^{(j)}x^{p-1} + b_2^{(j)}x^{p-2} + \dots + b_p^{(j)}\right) - \right.$$
$$\left. - \left(\frac{b_{p+1}^{(j)}}{x} + \frac{b_{p+2}^{(j)}}{x^2} + \dots + \frac{b_{p+q}^{(j)}}{x^q}\right)\right] \tag{3}$$

From (3) it is evident that the functions $Q_{2j-1}(x) - Q_{2j}(x)$ $(j = 1, 2, \dots, m)$ increase with increasing x. Hence, for $x > c_j > 0$, we have

$$Q_{2j-1}(x) - Q_{2j}(x) > Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0$$

whence for $x > R$ we get

$$P(x) = \sum_{j=1}^{m} [Q_{2j-1}(x) - Q_{2j}(x)] > 0$$

Thus all the positive roots x^+ of equation (2) of Sec. 5.2 satisfy the condition

$$x^+ \leqslant R$$

Example. Determine the bounds of the real roots of the equation

$$2x^5 - 100x^2 + 2x - 1 = 0 \tag{4}$$

Solution. Here $a_0 = 2$ and $A = \max(100, 2, 1) = 100$. Therefore the upper bound R of the positive roots of equation (4) is, by Theorem 2 of Sec. 5.1,

$$R = 1 + \frac{A}{a_0} = 1 + \frac{100}{2} = 51$$

Applying Lagrange's theorem and taking into account that

$$a_k = a_3 = -100 \quad \text{and} \quad B = \max(100, 1) = 100$$

we will get a much better estimate for the upper bound of the positive roots

$$R = 1 + \sqrt[3]{\frac{100}{2}} = 1 + \sqrt[3]{50} \approx 4.7$$

Finally, using the method of alternating sums, we find

$$2x^5 - 100x^2 = 2x^2(x^3 - 50) > 0$$

for $x > \sqrt[3]{50}$ (say, for $x > 3.7$) and

$$2x - 1 = 2\left(x - \frac{1}{2}\right) > 0 \text{ for } x > 0.5$$

Hence we can take

$$R = \max(3.7, 0.5) = 3.7$$

To determine the lower bound r of the positive roots of (4), set

$$x = \frac{1}{y}$$

Then equation (4) takes the form

$$y^5 - 2y^4 + 100y^3 - 2 = 0$$

We successively get

$$y^5 - 2y^4 = y^4(y - 2) > 0 \quad \text{for} \quad y > 2$$

and

$$100y^3 - 2 = 100(y^3 - 0.02) > 0 \quad \text{for} \quad y > 0.3$$

Consequently

$$R_1 = \max(2,\ 0.3) = 2$$

and

$$r = \frac{1}{R_1} = 0.5$$

To find the bound of the negative roots in equation (4) put

$$x = -z$$

Then

$$2z^5 + 10z^2 + 2z + 1 = 0 \tag{4'}$$

Since the coefficients of equation (4′) are positive or zero, this equation does not have any positive roots and so the given equation (4) does not have any negative roots.

5.4 NEWTON'S METHOD

Theorem (Newton). *If for* $x = c > 0$ *the polynomial* $P(x)$ *and all its derivatives* $P'(x), P''(x), \ldots, P^{(n)}(x)$ *are nonnegative:*

$$P^k(c) \geqslant 0 \qquad (k = 0,\ 1,\ 2,\ \ldots,\ n) \tag{1}$$

and $P^{(n)}(c) = n!\ a_0 > 0$, *then* $R = c$ *can be taken for the upper bound of the positive roots of the equation*

$$P(x) = 0 \tag{2}$$

Proof. With $x > c$, taking into account inequality (1), we have, on the basis of Taylor's formula,

$$P(x) = P(c) + P'(c)(x-c) + \ldots + \frac{P^{(n)}(c)}{n!}(x-c)^n > 0$$

Hence, all positive roots x^+ of equation (2) satisfy the inequality

$$x^+ \leqslant c$$

Note. In practical applications of Newton's theorem, the trial-and-error method is used (say, via the Horner scheme) to find a monotonic increasing sequence of positive numbers

$$0 < c_1 \leqslant c_2 \leqslant \ldots \leqslant c_{n-1} \leqslant c_n$$

for which the following inequalities hold true:

$$\begin{aligned} &P^{(n-1)}(c_1) \geqslant 0, \\ &P^{(n-2)}(c_2) \geqslant 0, \\ &\cdots\cdots\cdots \\ &P'(c_{n-1}) \geqslant 0, \\ &P(c_n) \geqslant 0 \end{aligned}$$

Such numbers definitely exist since for $a_0 > 0$ we have

$$P^{(m)}(x) \longrightarrow +\infty \qquad (m = 0, 1, 2, \ldots, n-1)$$

as $x \longrightarrow +\infty$. We can finally take $c = c_n$.

Indeed, since

$$P^{(n)}(x) = n!\, a_0 > 0$$

the function $P^{(n-1)}(x)$ is an increasing function and, hence, for $x > c_1$ we have

$$P^{(n-1)}(x) > P^{(n-1)}(c_1) \geqslant 0$$

From this inequality it follows that the function $P^{(n-2)}(x)$ is increasing in the interval $[c_1, +\infty)$ and therefore we get, for $x > c_2 \geqslant c_1$,

$$P^{(n-2)}(x) > P^{(n-2)}(c_2) \geqslant 0$$

Reasoning in this fashion consistently, we finally are assured that $P(x)$ is an increasing function in the interval $[c_{n-1}, +\infty)$ and hence for $x > c_n \geqslant c_{n-1}$ we have

$$P(x) > P(c_n) \geqslant 0$$

Which means that $x^+ \leqslant c_n$.

Example. Consider the equation (given in the example of Sec. 5.3)

$$P(x) = 2x^5 - 100x^2 + 2x - 1 = 0$$

Here

$$\begin{aligned} P'(x) &= 10x^4 - 200x + 2, \\ P''(x) &= 40x^3 - 200, \\ P'''(x) &= 120x^2, \\ P^{\mathrm{IV}}(x) &= 240x, \\ P^{\mathrm{V}}(x) &= 240 \end{aligned}$$

Obviously $P'''(x) > 0$, $P^{\mathrm{IV}}(x) > 0$, $P^{\mathrm{V}}(x) > 0$ for $x > 0$. We have

$$P''(x) = 40(x^3 - 5) > 0 \quad \text{for} \quad x \geqslant 2$$

We assume $c_1 = c_2 = c_3 = 2$. Since

$$P'(2) = 10 \cdot 16 - 200 \cdot 2 + 2 < 0$$

we determine the sign of the number

$$P'(3) = 10 \cdot 81 - 200 \cdot 3 + 2 > 0$$

We can take $c_4 = 3$. We have

$$P(3) = 2 \cdot 243 - 100 \cdot 9 + 2 \cdot 3 - 1 < 0$$

and so we compute

$$P(4)=2\cdot 1024-100\cdot 16+2\cdot 4-1>0$$

Thus, $c_5=4$, and the upper bound of the positive roots of the given equation is

$$R=4$$

The estimate via Newton's method is more exact than that given above on the basis of Lagrange's method, but less so than the estimate obtained by the method of alternating sums (see example of Sec. 5.3).

5.5 THE NUMBER OF REAL ROOTS OF A POLYNOMIAL

After the bounds have been established of the positive and negative roots of an algebraic equation

$$P(x)=0 \tag{1}$$

where $P(x)$ is a given polynomial, the question arises as to the number of real roots of the given equation on some known interval (a, b).

A general picture concerning the number of real roots of equation (1) on the interval (a, b) is given by the graph of the function $y=P(x)$ (Fig. 43), where the roots x_1, x_2, x_3 are found as the abscissas of the points of intersection of the graph with the x-axis.

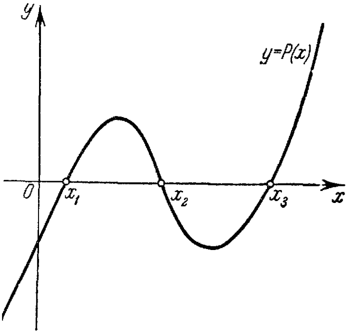


Fig. 43

We note the simple peculiarities of an integral polynomial.

(1) If $P(a)P(b)<0$, then on the interval (a, b) there is an odd number of roots of $P(x)$, counting multiplicities.

(2) If $P(a)P(b)>0$, then on the interval (a, b) there are no roots of the polynomial $P(x)$ or there is an even number of such roots.

The question of the number of real roots of an algebraic equation on a given interval is solved completely by the *Sturm method* [1], [2].

First let us introduce the notion of the number of sign-changes in a set of numbers.

Definition. *Suppose we have an ordered finite set of real numbers different from zero:*

$$c_1,\ c_2,\ \ldots,\ c_n \qquad (n\geqslant 2) \tag{2}$$

We say that there is a change of sign for a pair of two successive elements c_k, c_{k+1} *of* (2) *if these elements have opposite signs, that is,*

$$c_k c_{k+1} < 0$$

and there is no change of sign if the signs are the same; thus

$$c_k c_{k+1} > 0$$

The total number of changes of sign in all pairs of successive elements c_k, c_{k+1} $(k = 1, 2, \ldots, n-1)$ of (2) is called the *number of sign-changes* (*variations of sign*) in (2).

For a given polynomial $P(x)$, we form the *Sturm sequence*

$$P(x),\ P_1(x),\ P_2(x),\ \ldots,\ P_m(x) \tag{3}$$

where $P_1(x) = P'(x)$, $P_2(x)$ is the remainder, with reversed sign, left after the division of the polynomial $P(x)$ by $P_1(x)$, $P_3(x)$ is the remainder, with reversed sign, after the division of the polynomial $P_1(x)$ by $P_2(x)$, and so forth. The polynomials $P_k(x)$ $(k = 2, \ldots, m)$ may be found with the aid of a slightly modified Euclidean algorithm; if the polynomial $P(x)$ does not have any multiple roots, then the last element $P_m(x)$ in the Sturm sequence is a nonzero real number. Note that the elements in a Sturm sequence can be computed to within a positive numerical factor.

Denote by $N(c)$ the number of sign-changes in a Sturm sequence for $x = c$ provided that the zero elements of the sequence have been crossed out.

Sturm's theorem. *If a polynomial* $P(x)$ *does not have multiple roots and* $P(a) \neq 0$, $P(b) \neq 0$, *then the number of its real roots* $N(a, b)$ *on the interval* $a < x < b$ *is exactly equal to the number of lost sign-changes in the Sturm sequence of the polynomial* $P(x)$ *when going from* $x = a$ *to* $x = b$, *that is,*

$$N(a, b) = N(a) - N(b) \tag{4}$$

Corollary 1. If $P(0) \neq 0$, then the number N_+ of positive and the number N_- of negative roots of the polynomial $P(x)$ are respectively

$$N_+ = N(0) - N(+\infty)$$

and

$$N_- = N(-\infty) - N(0)$$

Corollary 2. For all roots of a polynomial $P(x)$ of degree n to be real, in the absence of multiple roots, it is necessary and sufficient that the following condition hold:

$$N(-\infty) - N(+\infty) = n$$

Thus, if

$$P(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

where $a_0 > 0$, then all roots of the equation $P(x) = 0$ will be real if and only if [1]: (1) the Sturm sequence has a maximum number of elements $n+1$, that is $m = n$, and (2) the inequalities $P_k(+\infty) > 0$ $(k = 1, 2, \ldots, n)$ hold true; thus the leading coefficients of all Sturm functions $P_k(x)$ must be positive.

Example. Determine the number of positive and negative roots of the equation

$$x^4 - 4x + 1 = 0 \tag{5}$$

Solution. The Sturm sequence is of the form

$$\begin{aligned} P(x) &= x^4 - 4x + 1, \\ P_1(x) &= x^3 - 1, \\ P_2(x) &= 3x - 1, \\ P_3(x) &= 1 \end{aligned}$$

whence

$$N(-\infty) = 2, \quad N(0) = 2, \quad N(+\infty) = 0$$

Hence, equation (5) has

$$N_+ = 2 - 0 = 2$$

positive roots and

$$N_- = 2 - 2 = 0$$

negative roots. And so two roots of (5) are complex roots.

We can isolate the roots of algebraic equations via the Sturm sequence by partitioning the interval (a, b) containing all real roots of the equation into a finite number of subintervals (α, β) such that

$$N(\alpha) - N(\beta) = 1$$

5.6 THE THEOREM OF BUDAN-FOURIER

Since the construction of a Sturm sequence generally involves unwieldy computations, for practical purposes one usually confines himself to simpler particular techniques for counting the number of real roots of algebraic equations.

Let us refine the counting of the number of variations of sign in a sequence of numbers.

Definition. *Suppose we have a finite ordered sequence of real numbers*

$$c_1, c_2, \ldots, c_n \tag{1}$$

where $c_1 \neq 0$ *and* $c_n \neq 0$.

On the one hand, we use the term lower number of variations of sign $\underline{N}$ *of the sequence* (1) *for the number of sign-changes in an appropriate subsequence that does not contain zero elements.*

On the other hand, we use the term upper number of variations of sign $\overline{N}$ *of a sequence of numbers* (1) *for the number of sign-changes in the transformed sequence* (1) *where the zero elements*

$$c_k = c_{k+1} = \ldots = c_{k+l-1} = 0$$

($c_{k-1} \neq 0$, $c_{k+l} \neq 0$) *are replaced by elements* $\tilde{c}_{k+i}$ ($i = 0, 1, 2, \ldots, l-1$) *such that*

$$\operatorname{sgn} \tilde{c}_{k+i} = (-1)^{l-i} \operatorname{sgn} c_{k+l} \tag{2}$$

It is obvious that if (1) has no zero elements, then the number N of sign-changes in the sequence coincides in meaning with its lower $\underline{N}$ and upper $\overline{N}$ numbers of variations of sign:

$$N = \underline{N} = \overline{N}$$

Generally speaking, $\overline{N} \geqslant \underline{N}$.

Example 1. Determine the lower number and the upper number of changes of sign in the sequence

$$1,\ 0,\ 0,\ -3,\ 1$$

Solution. Ignoring zeros, we have

$$\underline{N} = 2$$

To count $\overline{N}$ by formula (2), form the sequence

$$1,\ -\varepsilon,\ \varepsilon,\ -3,\ 1$$

where $\varepsilon > 0$, whence

$$\overline{N} = 4$$

Theorem (Budan-Fourier). *If the numbers* a *and* b ($a < b$) *are not roots of a polynomial* $P(x)$ *of degree* n, *then the number* $N(a, b)$ *of real roots of the equation*

$$P(x) = 0 \tag{3}$$

lying between a *and* b *is equal to the minimal number* ΔN *of sign-changes lost in the sequence of successive derivatives*

$$P(x),\ P'(x),\ \ldots,\ P^{(n-1)}(x),\ P^{(n)}(x) \tag{4}$$

when going from $x = a$ *to* $x = b$, *or less than* ΔN *by an even number:*

$$N(a, b) = \Delta N - 2k$$

where

$$\Delta N = \underline{N}(a) - \overline{N}(b)$$

and $\underline{N}(a)$ is the lower number of variations of sign in the sequence (4) for $x = a$, $\overline{N}(b)$ is the upper number of variations of sign in that sequence for $x=b\left[k=0,\ 1,\ \ldots,\ E\left(\frac{\Delta N}{2}\right)\right]$ (see [1]).

It is assumed here that each root of equation (3) is counted according to its multiplicity. If the derivatives $P^{(k)}(x)$ $(k=1,\ 2,\ \ldots,\ n)$ do not vanish at $x=a$ and $x=b$, then counting the signs is simplified, namely:

$$\Delta N=N(a)-N(b)$$

Corollary 1. If $\Delta N=0$, then there are no real roots of equation (3) between a and b.

Corollary 2. If $\Delta N=1$, then there is exactly one real root of equation (3) between a and b.

Note. To count the number of lost signs ΔN in sequence (4), we form two expansions using Horner's scheme:

$$P(a+h)=\alpha_0+\alpha_1 h+\alpha_2 h^2+\ldots+\alpha_n h^n \tag{5}$$

and

$$P(b+h)=\beta_0+\beta_1 h+\beta_2 h^2+\ldots+\beta_n h^n \tag{6}$$

Let $\underline{N}(a)$ be the lower number of variations of sign of the coefficients in expansion (5) and, respectively, $\overline{N}(b)$ the upper number of variations of sign of the coefficients in expansion (6). Since

$$\alpha_k=\frac{P^{(k)}(a)}{k!},\qquad \beta_k=\frac{P^{(k)}(b)}{k!}\qquad (k=0,\ 1,\ 2,\ \ldots,\ n)$$

it follows that the signs of the numbers α_k and β_k coincide with the signs of sequence (4) when $x=a$ and $x=b$. Therefore

$$\Delta N=\underline{N}(a)-\overline{N}(b)$$

Example 2. Determine the number of real roots of the equation

$$P(x)\equiv x^3-x^2+2x-3=0 \tag{7}$$

in the interval (0, 2).

Solution. Here $N(0)$ is clearly the number of variations of sign in the sequence of numbers

$$-3,\ 2,\ -1,\ 1$$

that is,

$$N(0)=3$$

The expansion of $P(2+h)$ is obtained by means of Horner's scheme:

$$\begin{array}{rrrrl}
1 & -1 & 2 & +3 & \underline{|2} \\
 & 2 & 2 & 8 & \\ \cline{1-4}
1 & 1 & 4 & \boxed{5} & \\
 & 2 & 6 & & \\ \cline{1-4}
1 & 3 & \boxed{10} & & \\
 & 2 & & & \\ \cline{1-4}
1 & \boxed{5} & & & \\
\boxed{1} & & & &
\end{array}$$

Hence, $N(2)$ is the number of variations of sign in the sequence of numbers

$$5,\ 10,\ 5,\ 1$$

or $N(2)=0$.

From this

$$\Delta N = N(0) - N(2) = 3$$

Thus, equation (7) has three real roots or one real root in the interval $(0, 2)$.

Descartes' rule of signs. *The number of positive roots of an algebraic equation*

$$P(x) \equiv a_0x^n + a_1x^{n-1} + \ldots + a_n = 0, \qquad (a_0 \neq 0) \tag{8}$$

a root of multiplicity m *being counted as* m *roots, is equal to the number of variations in sign in the sequence of coefficients*

$$a_0,\ a_1,\ a_2,\ \ldots,\ a_n \tag{9}$$

(where the coefficients equal to zero are not counted) *or less than that number by an even integer.*

Descartes' rule is an instance of the application of the Budan-Fourier theorem to the interval $(0, +\infty)$. Indeed, since

$$P^{(k)}(0) = k!\,a_{n-k} \qquad (k=0,\ 1,\ \ldots,\ n)$$

sequence (9) is, to within positive factors, a collection of derivatives $P^{(k)}(0)$ $(k=0, 1, 2, \ldots, n)$ written in descending order. Therefore, the number of variations in sign in the sequence (9) is equal to $\underline{N}(0)$, zero coefficients not being counted. On the other hand, the derivatives $P^{(k)}(+\infty)$ $(k=0,\ 1,\ 2,\ \ldots,\ n)$ clearly have one and the same sign and, hence, $\overline{N}(+\infty)=0$. We therefore have

$$\Delta N = \underline{N}(0) - \overline{N}(+\infty) = \underline{N}(0)$$

and, on the basis of the Budan-Fourier theorem, the number of positive roots of (8) is either equal to ΔN or is less than ΔN by an even integer.

Corollary. If the coefficients of equation (8) are different from zero, then the number of negative roots of (8) (counting multiplicities) is equal to the number of nonvariations of sign in the sequence (9) of its coefficients or is less than that number by an even integer. The proof of this assertion follows directly from the application of Descartes' rule to the polynomial $P(-x)$.

Also, let us give a necessary criterion for the real nature of all roots of a polynomial.

Hua's Theorem. *If an equation*

$$a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_n = 0 \tag{10}$$

has real coefficients and all its roots are real, then the square of each nonextreme coefficient of the equation is greater than the product of two adjacent coefficients; that is, we have the inequalities

$$a_k^2 > a_{k-1}a_{k+1} \quad (k = 1, 2, \dots, n-1)$$

Corollary. If for some k we have the inequality

$$a_k^2 \leqslant a_{k-1}a_{k+1}$$

then the equation (10) has at least one pair of complex roots.

Example 3. Determine the composition of the roots of the equation

$$x^4 + 8x^3 - 12x^2 + 104x - 20 = 0 \tag{11}$$

Solution. Since

$$(-12)^2 < 8 \cdot 104$$

it follows that equation (11) has complex roots and, hence, the number of real roots of the equation does not exceed two. In the sequence of coefficients of equation (11) there are $\Delta N = 3$ variations of sign and $\Delta P = 1$ nonvariation of sign. We thus conclude, on the basis of Descartes' rule and the corollary to it and taking into account the presence of complex roots, that equation (11) has one positive root, one negative root and a pair of complex roots.

5.7 THE UNDERLYING PRINCIPLE OF THE METHOD OF LOBACHEVSKY-GRAEFFE

Consider the nth degree algebraic equation

$$a_0x^n + a_1x^{n-1} + \dots + a_n = 0 \tag{1}$$

where $a_0 \neq 0$. Suppose that the roots $x_1, x_2, \dots, x_n$ of equation (1) are such that

$$|x_1| \gg |x_2| \gg |x_3| \gg \ldots \gg |x_n| \tag{2}$$

That is, the roots are distinct in modulus and the modulus of any one is much more than that of the following one.[1] In other words, we assume that the ratio of any two successive roots (in descending order) is a quantity small in modulus, i.e.

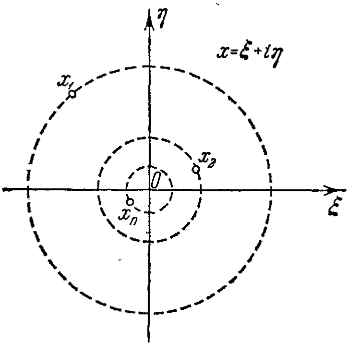


Fig. 44

$$\left.\begin{aligned} x_2 &= \varepsilon_1 x_1, \\ x_3 &= \varepsilon_2 x_2, \\ &\cdots\cdots \\ x_n &= \varepsilon_{n-1} x_{n-1} \end{aligned}\right\} \tag{3}$$

where $|\varepsilon_k| < \varepsilon$ and ε is a small quantity. For the sake of brevity we will call them *separated* roots (Fig. 44).

Now let us take advantage of the relations between the roots and coefficients of equation (1) (Sec. 5.1):

$$\left.\begin{aligned} x_1 + x_2 + \ldots + x_n &= -\frac{a_1}{a_0}, \\ x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n &= \frac{a_2}{a_0}, \\ \cdots\cdots\cdots\cdots\cdots&\cdots \\ x_1x_2\ldots x_n &= (-1)^n \frac{a_n}{a_0} \end{aligned}\right\}$$

From this, by virtue of the assumptions (3), we get

$$\left.\begin{aligned} x_1(1+E_1) &= -\frac{a_1}{a_0}, \\ x_1x_2(1+E_2) &= \frac{a_2}{a_0}, \\ \cdots\cdots\cdots&\cdots\cdots \\ x_1x_2\ldots x_n(1+E_n) &= (-1)^n \frac{a_n}{a_0} \end{aligned}\right\} \tag{4}$$

where $E_1, E_2, \ldots, E_n$ are quantities small in modulus compared to unity. Neglecting the quantities E_k $(k = 1, 2, \ldots, n)$ in (4), we get the approximate relations

---

[1] If the coefficients of (1) are real, then from Condition (2) it follows that all the roots of (1) are real.

$$\left.\begin{array}{l} x_1 \doteq -\dfrac{a_1}{a_0}, \\ x_1 x_2 = \dfrac{a_2}{a_0}, \\ \cdots\cdots\cdots\cdots \\ x_1 x_2 \ldots x_n = (-1)^n \dfrac{a_n}{a_0} \end{array}\right\} \qquad (5)$$

Whence we find the desired roots:

$$\left.\begin{array}{l} x_1 = -\dfrac{a_1}{a_0}, \\ x_2 = -\dfrac{a_2}{a_1}, \\ \cdots\cdots \\ x_n = -\dfrac{a_n}{a_{n-1}} \end{array}\right\} \qquad (6)$$

In other words, if the roots of equation (1) are separated, then they are determined approximately by the chain of linear equations

$$\begin{array}{c} a_0 x_1 + a_1 = 0, \\ a_1 x_2 + a_2 = 0, \\ \cdots\cdots \\ a_{n-1} x_n + a_n = 0 \end{array}$$

The accuracy of these roots depends on how small (in modulus) are the quantities ε_k in the relations (3).

To separate the roots, use (1) to obtain the transformed equation

$$a_0^{(m)} y^n + a_1^{(m)} y^{n-1} + \ldots + a_n^{(m)} = 0 \qquad (7)$$

whose roots $y_1, y_2, \ldots, y_n$ are the *m*th powers of the roots $x_1, x_2, \ldots, x_n$ of equation (1), that is

$$y_k = x_k^m \quad (k = 1, 2, \ldots, n) \qquad (8)$$

If the roots of equation (1), which we assume to be arranged in descending order of moduli, are distinct in modulus, the roots of equation (7) will be separated if m is a sufficiently high power, since

$$\frac{y_k}{y_{k-1}} = \left(\frac{x_k}{x_{k-1}}\right)^m \to 0 \quad \text{as} \quad m \to \infty$$

For example, let

$$x_1 = 2, \quad x_2 = 1.5, \quad x_3 = 1$$

For $m = 100$ we have

$$y_1 = 1.27 \cdot 10^{30}, \quad y_2 = 4.06 \cdot 10^{17}, \quad y_3 = 1$$

and, hence,

$$\frac{y_2}{y_1}=3.2\cdot 10^{-13}, \quad \frac{y_3}{y_2}=2.5\cdot 10^{-18}$$

Ordinarily, for the exponent m one takes a power of the number 2, that is, we put $m=2^p$, where p is a natural number; the transformation itself is executed in p steps, an equation being formed at each step (the roots of the equation are the squares of the roots of the preceding equation).

Approximating the roots $y_k (k=1, 2, \ldots, n)$ from formulas (8), we can determine the roots of the original equation (1). The accuracy of the computations depends on the smallness of the ratio of moduli of successive roots in the transformed equation.

The idea of this method for computing roots was suggested by Lobachevsky, and a practically convenient scheme of computation was advanced by Graeffe.

The advantage of the Lobachevsky-Graeffe method lies in the fact that it does not require the roots to be isolated. It is only necessary to get rid of multiple roots by the device given in Sec. 5.1. The actual computation of the roots is uniform and regular. As we will soon see, the method is also suitable for finding complex roots. One inconvenience of the method is that it involves large numbers. Another is the absence of a sufficiently reliable check on the computations, and there are difficulties in estimating the accuracy of the result obtained.

Note that if the roots of equation (1) are distinct but the moduli of some of them are nearly equal, the convergence of the Lobachevsky-Graeffe method is extremely slow. In this case, it is advisable to regard such roots as equal in modulus and to apply special computational techniques.

5.8 THE ROOT-SQUARING PROCESS

We will now show how one can easily set up an equation whose roots are the squares (taken with the minus sign) of the roots of the given algebraic equation. This latter procedure is used for reasons of convenience in order to avoid, as far as possible, the appearance of negative coefficients. The transition from roots $x_k (k=1, 2, \ldots, n)$ to the roots

$$y_k=-x_k^2 \tag{1}$$

will for brevity be called *root squaring*.

Let

$$P(x)\equiv a_0x^n+a_1x^{n-1}+\ldots+a_n=0$$

be the given equation, where $a_0\neq 0$.

Denoting the roots of this equation by $x_1, x_2, \ldots, x_n$ we have

$$P(x) = a_0(x - x_1)(x - x_2) \ldots (x - x_n)$$

whence

$$P(-x) = (-1)^n a_0 (x + x_1)(x + x_2) \ldots (x + x_n)$$

Consequently

$$P(x) P(-x) = (-1)^n a_0^2 (x^2 - x_1^2)(x^2 - x_2^2) \ldots (x^2 - x_n^2) \qquad (2)$$

Setting

$$y = -x^2$$

we get, by formula (2), the polynomial

$$Q(y) = P(x) P(-x)$$

whose roots are the numbers

$$y_k = -x_k^2 \quad (k = 1, 2, \ldots, n)$$

Since

$$P(-x) = (-1)^n [a_0 x^n - a_1 x^{n-1} + a_2 x^{n-2} - \ldots + (-1)^n a_n]$$

we have, after multiplying out the polynomials $P(x)$ and $P(-x)$,

$$P(x) P(-x) = (-1)^n [a_0^2 x^{2n} - (a_1^2 - 2a_0 a_2) x^{2n-2} + \\ + (a_2^2 - 2a_1 a_3 + 2a_0 a_4) x^{2n-4} - \ldots + (-1)^n a_n^2]$$

Hence, the equation we are interested in is

$$Q(y) \equiv A_0 y^n + A_1 y^{n-1} + A_2 y^{n-2} + \ldots + A_n = 0$$

where

$$\begin{aligned} A_0 &= a_0^2, \\ A_1 &= a_1^2 - 2a_0 a_2, \\ A_2 &= a_2^2 - 2a_1 a_3 + 2a_0 a_4, \\ &\ldots\ldots\ldots\ldots \\ A_n &= a_n^2 \end{aligned}$$

We write more compactly

$$A_k = a_k^2 + 2 \sum_{s=1}^{k} (-1)^s a_{k-s} a_{k+s} \quad (k = 0, 1, 2, \ldots, n)$$

where it is assumed that $a_s = 0$ for $s < 0$ and $s > n$.

Rule. *In root squaring, each coefficient of the transformed equation is equal to the square of the earlier coefficient minus twice the product of the adjacent coefficients plus twice the product of the next two coefficients, etc.* If the required coefficient is absent, it is considered equal to zero.

5.9 THE LOBACHEVSKY-GRAEFFE METHOD FOR THE CASE OF REAL AND DISTINCT ROOTS

Suppose the roots $x_1, x_2, \ldots, x_n$ of an nth degree equation with real coefficients

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

are real and unequal in modulus. Arrange them in order of decreasing moduli:

$$|x_1| > |x_2| > \ldots > |x_n|$$

Repeatedly applying the root-squaring process, we form the equation

$$b_0y^n + b_1y^{n-1} + \ldots + b_n = 0 \tag{2}$$

whose roots are the numbers

$$y_k = -x_k^{2^p} \qquad (k = 1, 2, \ldots, n) \tag{3}$$

If p is sufficiently great, the roots $y_1, y_2, \ldots, y_n$ are separated and, on the basis of the results of Sec. 5.7, can be determined from the chain of linear equations

$$\begin{gathered} b_0y_1 + b_1 = 0, \\ b_1y_2 + b_2 = 0, \\ \cdots\cdots\cdots \\ b_{n-1}y_n + b_n = 0 \end{gathered}$$

From this we get

$$x_k = \pm \sqrt[2^p]{-y_k} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \qquad (k = 1, 2, \ldots, n) \tag{4}$$

The signs of the roots x_k are determined by a rough guess, by substitution into the given equation, or on the basis of the relations between the roots and the coefficients of the equations. The process of root squaring usually continues until the doubled products cease to affect the first main terms of the coefficients of the transformed equation.

Rule. *The process of root squaring is terminated if the coefficients of some transformed equation are equal, to within the accuracy of the computations, to the squares of the corresponding coefficients of the preceding transformed equation due to the absence of doubled products.*

Indeed, if the transformed equation corresponding to the power 2^{p+1} has the form

$$c_0z^n + c_1z^{n-1} + \ldots + c_n = 0$$

and the relations

$$c_k = b_k^2 \qquad (k = 0, 1, 2, \ldots, n)$$

hold, then we clearly get

$$|x_k| = \sqrt[2^{p+1}]{\frac{c_k}{c_{k-1}}} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}}$$

Thus, under the circumstances we cannot improve the accuracy of the root computations.

Since, when applying the Lobachevsky-Graeffe method, the coefficients of the transformed equations generally grow rapidly, it is useful to isolate their orders by writing the coefficients in powers-of-ten notation $\alpha \cdot 10^m$, where $|\alpha| < 10$ and m is an integer. It is advisable to use logarithms in computations requiring extreme accuracy (see [5]).

Example. Use the Lobachevsky-Graeffe method to find the roots of the equation

$$x^3 - 3x + 1 = 0 \tag{5}$$

Solution. The results of the computations carried to four significant digits are tabulated in Table 8.

TABLE 8

COMPUTATION OF REAL ROOTS BY THE LOBACHEVSKY-GRAEFFE METHOD

Power	x^3	x^2	x	x^0
1	1	0	—3	1
		0	9	
		6	0	
2	1	6	9	1
		36	81	
		—18	—12	
4	1	18	69	1
		$3.24 \cdot 10^2$	$4.761 \cdot 10^3$	
		$-1.38 \cdot 10^2$	$-0.036 \cdot 10^3$	
8	1	$1.86 \cdot 10^2$	$4.725 \cdot 10^3$	1
		$3.460 \cdot 10^4$	$2.233 \cdot 10^7$	
		$-0.945 \cdot 10^4$	0	
16	1	$2.515 \cdot 10^4$	$2.233 \cdot 10^7$	1
		$6.325 \cdot 10^8$	$4.986 \cdot 10^{14}$	
		$-0.447 \cdot 10^8$	0	
32	1	$5.878 \cdot 10^8$	$4.986 \cdot 10^{14}$	1
		$3.455 \cdot 10^{17}$	$2.486 \cdot 10^{29}$	
		$-0.010 \cdot 10^{17}$	0	
64	1	$3.445 \cdot 10^{17}$	$2.486 \cdot 10^{29}$	1
		$1.187 \cdot 10^{35}$	$6.180 \cdot 10^{58}$	
		0	0	
128	1	$1.187 \cdot 10^{35}$	$6.180 \cdot 10^{58}$	1

Stopping with the 64th power of the roots, we have

$$-x_1^{64}+3.445\cdot 10^{17}=0,$$
$$-3.445\cdot 10^{17}\cdot x_2^{64}+2.486\cdot 10^{29}=0,$$
$$-2.486\cdot 10^{29}\cdot x_3^{64}+1=0$$

whence

$$x_1=\pm\sqrt[64]{3.445\cdot 10^{17}},$$
$$x_2=\pm\sqrt[64]{\frac{2.486}{3.445}\cdot 10^{12}},$$
$$x_3=\pm\sqrt[64]{\frac{1}{2.486}\cdot 10^{-29}}$$

Taking logarithms we obtain

$$\log_{10}|x_1|=\frac{1}{64}\cdot 17.53719=0.27402,$$
$$\log_{10}|x_2|=\frac{1}{64}\cdot 11.85831=0.18528;$$
$$\log_{10}|x_3|=\frac{1}{64}\cdot(-29.39550)=\bar{1}.54070$$

and, consequently,

$$x_1=\pm 1.879,$$
$$x_2=\pm 1.532,$$
$$x_3=\pm 0.347$$

In determining the signs of the roots, note that by Descartes' rule, equation (5) has one negative root and two positive roots,[1] and

$$x_1+x_2+x_3=0 \tag{6}$$

Therefore, the negative root must be the largest in modulus and we finally get

$$x_1=-1.879,$$
$$x_2=1.532,$$
$$x_3=0\ 347$$

Relation (6) holds true within the specified accuracy. By way of comparison, we give the exact values of the roots obtained by Cardan's formula:

$$x_1=2\cos 160^\circ=-1.87938,$$
$$x_2=2\cos 40^\circ\ =1.53208,$$
$$x_3=2\cos 80^\circ\ =0.34730$$

---

[1] We take into account the fact that the equation $P(x)\equiv x^3-3x+1=0$ has positive roots since $P(0)>0$ and $P(1)<0$.

It will be noted that in our case the computation of roots was somewhat simplified because the extreme coefficients of the equation were equal to unity. Generally speaking, when using the Lobachevsky-Graeffe method it is advisable first to transform the equation so that the leading coefficient is equal to unity and the constant term is equal to ±1 (see [5]).

5.10 THE LOBACHEVSKY-GRAEFFE METHOD FOR THE CASE OF COMPLEX ROOTS

Let us now generalize the concept of separation of roots.

Suppose the roots $x_1, x_2, \ldots, x_n$ of the equation

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

satisfy the conditions

$$|x_1| \geqslant |x_2| \geqslant \ldots \geqslant |x_m| \gg |x_{m+1}| \geqslant |x_{m+2}| \geqslant \ldots \geqslant |x_n| \tag{2}$$

In other words, it is assumed that the roots of equation (1) can be divided into two *categories* (groups):

$$x_1, x_2, \ldots, x_m \quad (m < n)$$

and

$$x_{m+1}, x_{m+2}, \ldots, x_n$$

so that the moduli of the roots of the first category are very great compared with those of the second category (see Fig. 45, where the roots lie in the cross-lined regions, while the interior of the nonlined annulus is free of roots and constitutes a "desert area").

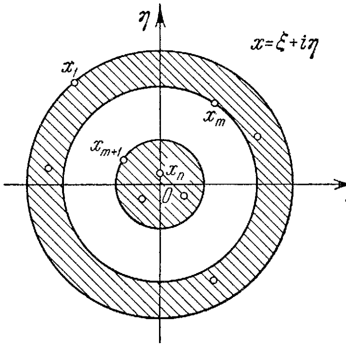


Fig. 45

Write down the first m relations between the roots and the coefficients of (1):

$$x_1 + x_2 + \ldots + x_m + (x_{m+1} + \ldots + x_n) = -\frac{a_1}{a_0},$$

$$x_1x_2 + x_2x_3 + \ldots + x_{m-1}x_m + (x_mx_{m+1} + \ldots + x_{n-1}x_n) = \frac{a_2}{a_0},$$

. . . .

$$x_1x_2\ldots x_m + (x_1x_2\ldots x_{m-1}x_{m+1} + \ldots + x_{n-m+1}x_{n-m+2}\ldots x_n) =$$
$$= (-1)^m \frac{a_m}{a_0}$$

Neglecting the terms with relatively small moduli (they are in brackets), we get the approximate relations

$$\left.\begin{aligned} &x_1+x_2+\ldots+x_m=-\frac{a_1}{a_0}, \\ &x_1x_2+x_2x_3+\ldots+x_{m-1}x_m=\frac{a_2}{a_0}, \\ &\cdots\cdots\cdots\cdots\cdots \\ &x_1x_2\ldots x_m=(-1)^m\frac{a_m}{a_0} \end{aligned}\right\} \qquad (3)$$

From this it follows that the roots $x_1, x_2, \ldots, x_m$ of the first category (with large moduli) are approximately the roots of the equation

$$a_0x^m+a_1x^{m-1}+\ldots+a_m=0 \qquad (4)$$

Of the remaining unused $n-m$ relations between the roots and coefficients of equation (1), we get

$$\begin{aligned} &x_1x_2\ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)+ \\ &\quad +x_2x_3\ldots x_{m+1}x_{m+2}+\ldots+x_{n-m}x_{n-m+1}\ldots x_n=(-1)^{m+1}\frac{a_{m+1}}{a_0}, \\ &x_1x_2\ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)+x_2x_3\ldots x_{m+1}x_{m+2}x_{m+3}+\ldots \\ &\quad \ldots+x_{n-m-1}\ldots x_n=(-1)^{m+2}\frac{a_{m+2}}{a_0}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &x_1x_2\ldots x_mx_{m+1}\ldots x_n=(-1)^n\frac{a_n}{a_0} \end{aligned}$$

Dropping the relatively small (in modulus) terms, we get the approximate relations

$$\begin{aligned} &x_1x_2\ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)=(-1)^{m+1}\frac{a_{m+1}}{a_0}, \\ &x_1x_2\ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)=(-1)^{m+2}\frac{a_{m+2}}{a_0}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &x_1x_2\ldots x_mx_{m+1}\ldots x_n=(-1)^n\frac{a_n}{a_0} \end{aligned}$$

Whence, using the last relation in formulas (3), we find

$$\left.\begin{aligned} &x_{m+2}+x_{m+3}+\ldots+x_n=-\frac{a_{m+1}}{a_m}, \\ &x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n=\frac{a_{m+2}}{a_m}, \\ &\cdots\cdots\cdots\cdots\cdots \\ &x_{m+1}x_{m+2}\ldots x_n=(-1)^{n-m}\frac{a_n}{a_m} \end{aligned}\right\} \qquad (5)$$

Consequently, the roots $x_{m+1}, x_{m+2}, \ldots, x_n$ of second category (with small moduli) are approximately the roots of the equation

$$a_m x^{n-m} + a_{m-1} x^{n-m-1} + \ldots + a_n = 0 \tag{6}$$

Thus, under our conditions, equation (1) decomposes into two equations of lower degree, each one of which approximately determines the roots belonging to one of the categories.

Arguing by analogy, we conclude that if the roots of (1) can be split into p categories

$$\begin{gathered} x_1, x_2, \ldots, x_{m_1}, \\ x_{m_1+1}, x_{m_1+2}, \ldots, x_{m_2}, \\ \cdots\cdots\cdots \\ x_{m_{p-1}+1}, x_{m_{p-1}+2}, \ldots, x_{m_p} \\ (m_1 + m_2 + \ldots + m_p = n) \end{gathered}$$

so that the condition

$$\begin{aligned} |x_1| \geqslant |x_2| \geqslant \ldots \geqslant |x_{m_1}| \gg |x_{m_1+1}| \geqslant |x_{m_1+2}| \geqslant \ldots \geqslant |x_{m_2}| \gg \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \gg |x_{m_{p-1}+1}| \geqslant |x_{m_{p-1}+2}| \geqslant \ldots \geqslant |x_{m_p}| \end{aligned}$$

holds, that is, the moduli of the roots belonging to lower categories considerably exceed the moduli of roots of higher categories (we may say that these roots are *separated in the group sense*), then the roots of each category can be determined in approximate fashion from the corresponding equations

$$\left.\begin{aligned} & a_0 x^{m_1} + a_1 x^{m_1-1} + \ldots + a_{m_1} = 0, \\ & a_{m_1} x^{m_2} + a_{m_1+1} x^{m_2-1} + \ldots + a_{m_1+m_2} = 0, \\ & \cdots\cdots\cdots\cdots\cdots\cdots \\ & a_{m_1+m_2+\ldots+m_{p-1}} x^{m_p} + a_{m_1+m_2+\ldots+m_{p-1}+1} x^{m_p-1} + \ldots \\ & \qquad \ldots + a_{m_1+m_2+\ldots+m_p} = 0 \end{aligned}\right\} \tag{7}$$

the powers of which are $m_1, m_2, \ldots, m_p$, respectively. In particular, if the roots of (1) are completely separated, then equations (7) are linear equations; a pair of complex roots, in the absence of other roots of the same modulus, will be associated with a quadratic equation in (7).

We consider here only the simplest cases when equation (1), whose coefficients are considered to be real, has one pair of complex roots or two pairs of complex roots with distinct moduli, the moduli of the real roots being distinct and different from the moduli of the complex roots. More general cases are discussed in Krylov [5] and Scarborough [6].

5.11 THE CASE OF A PAIR OF COMPLEX ROOTS

Suppose

$$\left.\begin{aligned} x_m &= u+iv, \\ x_{m+1} &= u-iv \end{aligned}\right\} \tag{1}$$

(u and v real, $v \neq 0$) are complex roots of equation (1) of Sec. 5.10, and all the other roots x_k ($k \neq m$, $k \neq m+1$) of this equation are real and satisfy the condition

$$|x_1| > |x_2| > \ldots > |x_m| = |x_{m+1}| > \ldots > |x_n| \tag{2}$$

Applying the root-squaring process, form the equation

$$b_0 y^n + b_1 y^{n-1} + \ldots + b_n = 0$$

whose roots are

$$y_k = -x_k^{2^p} \qquad (k = 1, 2, \ldots, n)$$

Given a sufficiently large natural p, the real roots $y_1, \ldots, y_{m-1}$, $y_{m+2}, \ldots, y_n$ will be separated with a high degree of accuracy and may be determined from the linear equations

$$\begin{gathered} b_0 y_1 + b_1 = 0, \\ \cdots\cdots\cdots\cdots \\ b_{m-2} y_{m-1} + b_{m-1} = 0, \\ b_{m+1} y_{m+2} + b_{m+2} = 0, \\ \cdots\cdots\cdots\cdots \\ b_{n-1} y_n + b_n = 0 \end{gathered}$$

whence we get

$$x_k = \pm \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \qquad (k \neq m, \ k \neq m+1)$$

The root-squaring process is terminated when the doubled products in the coefficients $b_1, \ldots, b_{m-1}, b_{m+1}, \ldots, b_n$ vanish in the next step (to within the specified accuracy). As for the coefficient b_m, it does not, generally speaking, include vanishing doubled products. It may even happen that these products dominate the square and the coefficient b_m has a changing sign. This is a characteristic sign of the presence of complex roots or roots with equal moduli in equation (1) of Sec. 5.10, and the unusual behaviour of the coefficient b_m indicates the **position** of such roots in the sequence (2) of moduli.

Note that the equation under consideration definitely has complex roots if the coefficient b_m changes sign; thus, in the case of real roots only, all the coefficients of the transformed equations will clearly be nonnegative.

According to this general theory, the roots y_m and y_{m+1}, which correspond to the complex roots x_m and x_{m+1}, approximately satisfy the quadratic equation

$$b_{m-1}y^2+b_my+b_{m+1}=0$$

Note that the coefficient b_m is the **middle** one. Since

$$x_kx_{k+1}=u^2+v^2=r^2$$

where

$$r=|x_k|=|x_{k+1}|$$

is the common modulus of the complex roots, and

$$y_my_{m+1}=x_k^{2^p}\cdot x_{k+1}^{2^p}=(x_kx_{k+1})^{2^p}=(r^2)^{2^p}$$

then, by the property of the roots of a quadratic equation, we have

$$(r^2)^{2^p}=\frac{b_{m+1}}{b_{m-1}}$$

And from this we determine the square of the modulus of the complex roots:

$$r^2=\sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}} \tag{3}$$

The easiest way to find the real part u of the complex roots is to use the relation

$$x_1+x_2+\ldots+x_{m-1}+(x_m+x_{m+1})+x_{m+2}+\ldots+x_n=-\frac{a_1}{a_0}$$

whence

$$2u=x_m+x_{m+1}=-\frac{a_1}{a_0}-\sum_{\substack{k\neq m\\ k\neq m+1}}'' x_k$$

and, consequently,

$$u=-\frac{a_1}{2a_0}-\frac{1}{2}\sum_{\substack{k\neq m\\ k\neq m+1}}'' x_k \tag{4}$$

Knowing, by virtue of formula (3), the common modulus r of the complex roots, we find the coefficient v of their imaginary part:

$$v=\sqrt{r^2-u^2} \tag{5}$$

Using formulas (4) and (5), determine the desired complex roots:

$$x_{m,\,m+1}=u\pm iv$$

It is also possible to seek complex roots in trigonometric form:

$$x_{m,\,m+1}=r(\cos\varphi\pm i\sin\varphi)$$

Example. Find the roots of the equation [7]

$$x^4+x^3-10x^2-34x-26=0 \tag{6}$$

Solution. The results of the computations, to four significant digits, are given in Table 9.

TABLE 9

COMPUTING COMPLEX ROOTS BY THE LOBACHEVSKY-GRAEFFE METHOD

Power	x^4	x^3	x^2	x	x^0
1	1	1 1 20	-10 100 68 -52	-34 1156 -520	-26
2	1	21 441 -239	116 $1.346\cdot10^{4}$ $-2.671\cdot10^{4}$ $0.135\cdot10^{4}$	636 $4.045\cdot10^{5}$ $-1.568\cdot10^{5}$	676
4	1	209 $4.368\cdot10^{4}$ $2.380\cdot10^{4}$	$-1.190\cdot10^{4}$ $1.416\cdot10^{8}$ $-1.035\cdot10^{8}$ $0.009\cdot10^{8}$	$2.477\cdot10^{5}$ $6.135\cdot10^{10}$ $1.088\cdot10^{10}$	$4.570\cdot10^{5}$
8	1	$6.748\cdot10^{4}$ $4.554\cdot10^{9}$ $-0.078\cdot10^{9}$	$3.90\cdot10^{7}$ $1.521\cdot10^{15}$ $-9.748\cdot10^{15}$ 0	$7.223\cdot10^{10}$ $5.216\cdot10^{21}$ $-0.016\cdot10^{21}$	$2.088\cdot10^{4}$
16	1	$4.476\cdot10^{9}$ $2.003\cdot10^{19}$ $0.002\cdot10^{19}$	$-8.227\cdot10^{15}$ $6.768\cdot10^{31}$ $-4.655\cdot10^{31}$ 0	$5.200\cdot10^{21}$ $2.704\cdot10^{43}$ 0	$4.360\cdot10^{22}$
32	1	$2.005\cdot10^{19}$ $4.020\cdot10^{38}$ 0	$2.113\cdot10^{31}$ $4.465\cdot10^{62}$ $-1.084\cdot10^{63}$ 0	$2.704\cdot10^{43}$ $7.312\cdot10^{86}$ 0	$1.901\cdot10^{45}$
64	1	$4.020\cdot10^{38}$	$-6.38\cdot10^{62}$	$7.312\cdot10^{86}$	$3.614\cdot10^{90}$

From Table 9 it is clear that in the fifth transformed equation (with 32nd powers of the roots, $2^5=32$), the real roots x_1 and x_4 (in descending order of moduli) are separated. These roots can be found from the two-term equations

$$-x_1^{32}+2.005\cdot10^{19}=0,$$
$$-2.704\cdot10^{43}x_4^{32}+1.901\cdot10^{45}=0$$

whence

$$x_1 = \pm \sqrt[32]{2.005 \cdot 10^{19}}, \qquad x_4 = \pm \sqrt[32]{\frac{1.901}{2.704} \cdot 10^2}$$

Taking logarithms, we have

$$\log_{10} |x_1| = \frac{1}{32} \cdot 19.30211 = 0.60319,$$

$$\log_{10} |x_4| = \frac{1}{32} \cdot (2.27898 - 0.43201) = 0.05772$$

Hence

$$x_1 = \pm 4.010, \qquad x_4 = \pm 1.142$$

A rough guess convinces us that the root x_1 is positive and the root x_4 is negative. We thus finally get

$$x_1 = 4.010, \qquad x_4 = -1.142$$

Since the transformed coefficient of x^2 changes sign, the given equation has complex roots $x = x_2$ and $x = x_3$ which are found from the three-term equation

$$2.005 \cdot 10^{19} y^2 + 2.113 \cdot 10^{31} y + 2.704 \cdot 10^{43} = 0$$

where

$$y = -x^{32}$$

By the general theory, the modulus of the roots

$$r = |x_2| = |x_3|$$

is found from formula (3):

$$r^2 = \sqrt[32]{\frac{2.704}{2.005} \cdot 10^{24}}$$

whence

$$\log_{10} r^2 = \frac{1}{32} \cdot (24.43201 - 0.30211) = 0.75406$$

and, therefore,

$$r^2 = 5.6763$$

Setting

$$x_2 = u + iv, \quad x_3 = u - iv$$

we get from

$$x_1 + x_2 + x_3 + x_4 = -1$$

the relation

$$u = \frac{1}{2}(-1 - 4.010 + 1.142) = -1.934$$

The coefficient of the imaginary part v is found from the formula

$$v = \sqrt{r^2 - u^2} = \sqrt{5.6763 - 3.7404} = \sqrt{1.9359} = 1.395$$

Hence

$$x_{2,3} = -1.934 \pm 1.395i$$

Note that the roots x_2 and x_3 may also be found from the relations between the roots and the coefficients of equation (6); namely, we have

$$x_1 + x_2 + x_3 + x_4 = -1,$$
$$x_1 x_2 x_3 x_4 = -26$$

and from this, using the values of x_1 and x_4 found above, we get

$$x_2 + x_3 = -3.869,$$
$$x_2 x_3 = 5.677$$

And so x_2 and x_3 may be found as the roots of the quadratic equation

$$x^2 + 3.869x + 5.677 = 0$$

whose solution yields

$$x_{2,3} = -1.934 \pm 1.391i$$

5.12 THE CASE OF TWO PAIRS OF COMPLEX ROOTS

Suppose equation (1) of Sec. 5.10 admits two pairs of complex roots:

$$x_k = u_1 + iv_1, \quad x_{k+1} = u_1 - iv_1$$

and

$$x_m = u_2 + iv_2, \quad x_{m+1} = u_2 - iv_2$$

with distinct moduli (u_1, v_1, u_2, v_2 real and $v_1 \neq 0$, $v_2 \neq 0$); all other roots $x_j (j \neq k,\ j \neq k+1,\ j \neq m,\ j \neq m+1)$ of this equation are real, distinct in absolute value, nonzero (zero roots can be isolated beforehand), and different from the complex roots in modulus; that is,

$$|x_1| > |x_2| > \ldots > |x_{k-1}| > |x_k| = |x_{k+1}| > \ldots > |x_m| =$$
$$= |x_{m+1}| > \ldots > |x_n| > 0 \quad (1)$$

As usual, performing the root-squaring process in the equation at hand up to some power 2^p, we get the transformed equation

$$b_0 y^n + b_1 y^{n-1} + \ldots + b_n = 0$$

whose roots are the numbers

$$y_j = -x_j^{2^p} \qquad (j = 1, 2, \ldots, n)$$

For a sufficiently large p, it will be seen that in passing to the power 2^{p+1} some of the coefficients c_j of the newly transformed equation

$$c_0 z^n + c_1 z^{n-1} + \ldots + c_n = 0$$

will be (to within the specified accuracy) squares of the corresponding coefficients b_j of the preceding transformed equation. Under our assumption (1), we finally get

$c_j = b_j^2$ for $j = 0, 1, 2, \ldots, k-1, k+1, \ldots, m-1, m+1, \ldots, n$

and

$$c_k \neq b_k^2 \quad \text{and} \quad c_m \neq b_m^2$$

This enables us to establish the position of the complex roots. Note that a change in the signs of the coefficients b_k and b_m for different exponents 2^p serves as a sufficient criterion of the presence of two pairs of complex roots of equation (1) of Sec. 5.10.

The real roots x_j of the equation under consideration are determined from the two-term equations

$$-b_{j-1} x_j^{2^p} + b_j = 0 \tag{2}$$

whence

$$x_j = \pm \sqrt[2^p]{\frac{b_j}{b_{j-1}}} \quad (j \neq k,\ j \neq k+1,\ j \neq m,\ j \neq m+1)$$

The complex roots x_k, x_{k+1} and x_m, x_{m+1} are found respectively from the three-term equations

$$b_{k-1} x^{2^{p+1}} - b_k x^{2^p} + b_{k+1} = 0 \tag{2$'$}$$

and

$$b_{m-1} x^{2^{p+1}} - b_m x^{2^p} + b_{m+1} = 0 \tag{2$''$}$$

Let us introduce the notation

$$r_1 = |x_k| = |x_{k+1}|$$

and

$$r_2 = |x_m| = |x_{m+1}|$$

Noting that

$$r_1^2 = x_k x_{k+1}$$

and

$$r_2^2 = x_m x_{m+1}$$

we can compute from equations (2′) and (2″) the squares of the moduli of the complex roots:

$$r_1^2 = \sqrt[2^p]{\frac{b_{k+1}}{b_{k-1}}} \quad \text{and} \quad r_2^2 = \sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}}$$

To determine the real parts u_1 and u_2 of the complex roots, use the relations between the roots and coefficients of equation (1) of Sec. 5.10. We have

$$x_2x_3\ldots x_n+x_1x_3\ldots x_n+\ldots+x_1x_2\ldots x_{n-1}=(-1)^{n-1}\frac{a_{n-1}}{a_0}$$

and

$$x_1x_2\ldots x_n=(-1)^n\frac{a_n}{a_0}$$

Dividing the first equation by the second, we get

$$\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}=-\frac{a_{n-1}}{a_n}$$

Besides,

$$x_1+x_2+\ldots+x_n=-\frac{a_1}{a_0}$$

whence, taking into account the relations

$$x_k+x_{k+1}+x_m+x_{m+1}=2u_1+2u_2$$

and

$$\frac{1}{x_k}+\frac{1}{x_{k+1}}+\frac{1}{x_m}+\frac{1}{x_{m+1}}=\frac{2u_1}{r_1^2}+\frac{2u_2}{r_2^2}$$

we have the following linear system of equations:

$$\left.\begin{aligned} u_1+u_2&=-\frac{a_1}{2a_0}-\frac{1}{2}\sigma,\\ \frac{u_1}{r_1^2}+\frac{u_2}{r_2^2}&=-\frac{a_{n-1}}{2a_n}-\frac{1}{2}\sigma' \end{aligned}\right\} \qquad (3)$$

where σ is the sum of the real roots and σ' is the sum of the reciprocals of the real roots:

$$\sigma=\sum_{j\neq k,\,k+1,\,m,\,m+1} x_j$$

and

$$\sigma'=\sum_{j\neq k,\,k+1,\,m,\,m+1}\frac{1}{x_j}$$

Finding u_1 and u_2 from system (3), we determine the coefficients v_1 and v_2 of the imaginary parts of the complex roots from the formulas

$$v_1=\sqrt{r_1^2-u_1^2},\qquad v_2=\sqrt{r_2^2-u_2^2}$$

We finally get

$$x_{k,\,k+1}=u_1\pm iv_1$$

and

$$x_{m,\,m+1}=u_2\pm iv_2$$

Example. Using the Lobachevsky-Graeffe method, solve the equation [7]

$$x^4+4x^2-3x+3=0 \tag{4}$$

Solution. Applying the root-squaring process up to the 16th power and carrying the result to four significant digits, we obtain the results as given in Table 10.

TABLE 10

COMPUTING TWO PAIRS OF COMPLEX ROOTS BY THE LOBACHEVSKY-GRAEFFE METHOD

Power	x^4	x^3	x^2	x	x^0
1	1	0 0 —8	4 16 0 6	—3 9 —24	3
2	1	—8 64 —44	22 484 —240 18	—15 225 —396	9
4	1	20 $4\cdot10^2$ $-5.24\cdot10^2$	262 $6.864\cdot10^4$ $0.684\cdot10^4$ $0.016\cdot10^4$	—171 $2.924\cdot10^4$ $-4.244\cdot10^4$	81
8	1	$-1.24\cdot10^2$ $1.538\cdot10^4$ $-15.128\cdot10^4$	$7.564\cdot10^4$ $5.723\cdot10^9$ $-0.003\cdot10^9$ 0	$-1.320\cdot10^4$ $1.743\cdot10^8$ $-9.927\cdot10^8$	$6.561\cdot10^3$
16	1	$-1.359\cdot10^5$	$5.720\cdot10^9$	$-8.184\cdot10^8$	$4.305\cdot10^7$

It is readily seen that in the next transformation the middle coefficient will be equal to the square of the earlier value, and so we stop the root-squaring process. Since for the 16th power there are two negative coefficients among the coefficients of the transformed equation, equation (4) admits two pairs of complex roots:

$$x_{1,\,2}=u_1\pm iv_1$$

and

$$x_{3,\,4}=u_2\pm iv_2$$

which, respectively, satisfy the three-term equations

$$x^{32}+1.359\cdot10^5\cdot x^{16}+5.720\cdot$$

and

$$5.720\cdot 10^9\cdot x^{32}+8.184\cdot 10^8\cdot x^{16}+4.305\cdot 10^7=0$$

From this we determine the squares of the moduli of these roots:

$$r_1^2=\sqrt[16]{5.720\cdot 10^9}=4.072$$

and

$$r_2^2=\sqrt[16]{\frac{4.305}{5.720}\cdot 10^{-2}}=0.7367$$

Since

$$\frac{1}{r_1^2}=0.2456,\ \frac{1}{r_2^2}=1.3574$$

it follows, on the basis of system (3), that to find the real parts u_1 and u_2 we have the system

$$u_1+u_2=0,$$
$$0.2456u_1+1.3574u_2=0.5$$

whence

$$u_1=-0.4497,$$
$$u_2=0.4497$$

Using the squares r_1^2 and r_2^2 of the moduli of the roots, we determine the coefficients v_1 and v_2 of the imaginary parts of the roots:

$$v_1=\sqrt{r_1^2-u_1^2}=1.967,$$
$$v_2=\sqrt{r_2^2-u_2^2}=0.731$$

Thus, the roots of equation (4) are of the form

$$x_{1,\,2}=-0.450\pm 1.967i$$

and

$$x_{3,\,4}=0.450\pm 0.731i$$

5.13 BERNOULLI'S METHOD

Suppose we have an algebraic equation

$$a_0x^n+a_1x^{n-1}+\ldots+a_n=0 \qquad (a_0\neq 0) \tag{1}$$

whose roots $x_1, x_2, \ldots, x_n$ are distinct.

On the basis of the coefficients a_k ($k=0, 1, \ldots, n$) we construct a so-called *difference equation*

$$a_0y_{n+i}+a_1y_{n+i-1}+\ldots+a_ny_i=0 \qquad (i=0, 1, 2, \ldots) \tag{2}$$

which is a recurrence relation relating $n+1$ arbitrary successive terms of the nonterminating sequence

$$y_0, y_1, y_2, \ldots, y_i, \ldots \tag{3}$$

The sequence (3) $y_i = f(i)$ $(i = 0, 1, 2, \ldots)$ whose terms satisfy the difference equation (2) is called the *solution* of the equation. To construct a solution y_i, it is sufficient to specify n *initial values* $y_0, y_1, \ldots, y_{n-1}$; the remaining terms $y_n, y_{n+1}, \ldots$ can be found in a step-by-step manner from equation (2).

Proof is given [8] in the theory of finite differences that if the roots $x_1, x_2, \ldots, x_n$ of an algebraic equation (1) are distinct, then any solution of the difference equation (2) is of the form

$$y_i = C_1x_1^i + C_2x_2^i + \ldots + C_nx_n^i \qquad (i = 0, 1, 2, \ldots) \tag{4}$$

where $C_1, C_2, \ldots, C_n$ are arbitrary constants. Thus, equation (1) is the *characteristic* equation of (2). The constants $C_1, C_2, \ldots, C_n$ can be found from the initial conditions:

$$\left.\begin{aligned} y_0 &= C_1 + C_2 + \ldots + C_n, \\ y_1 &= C_1x_1 + C_2x_2 + \ldots + C_nx_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{n-1} &= C_1x_1^{n-1} + C_2x_2^{n-1} + \ldots + C_nx_n^{n-1} \end{aligned}\right\} \tag{5}$$

Theorem. *Let the algebraic equation* (1) *have a unique maximum-modulus root* x_1. *Then the ratio of two successive terms* y_{i+1} *and* y_i *of the solution of the difference equation* (2) *tends (generally speaking) to a limit equal to* x_1; that is,

$$\lim_{i\to\infty} \frac{y_{i+1}}{y_i} = x_1 \tag{6}$$

Proof. Let

$$|x_1| > |x_2| \geqslant \ldots \geqslant |x_n| \tag{7}$$

Assuming the roots x_k $(k = 1, 2, \ldots, n)$ to be distinct, we get from formula (4)

$$y_i = x_1^i \left[C_1 + C_2\left(\frac{x_2}{x_1}\right)^i + \ldots + C_n\left(\frac{x_n}{x_1}\right)^i\right]$$

and

$$y_{i+1} = x_1^{i+1} \left[C_1 + C_2\left(\frac{x_2}{x_1}\right)^{i+1} + \ldots + C_n\left(\frac{x_n}{x_1}\right)^{i+1}\right]$$

whence

$$\frac{y_{i+1}}{y_i} = x_1 \cdot \frac{C_1 + C_2\left(\frac{x_2}{x_1}\right)^{i+1} + \ldots + C_n\left(\frac{x_n}{x_1}\right)^{i+1}}{C_1 + C_2\left(\frac{x_2}{x_1}\right)^{i} + \ldots + C_n\left(\frac{x_n}{x_1}\right)^{i}} \tag{8}$$

If $C_1 \neq 0$, then, passing to the limit in (8) as $i \to \infty$ and noting that by virtue of inequality (7) the limit relations

$$\left(\frac{x_2}{x_1}\right)^i \to 0, \quad \ldots, \quad \left(\frac{x_n}{x_1}\right)^i \to 0$$

hold, we will have

$$\lim_{i\to\infty} \frac{y_{i+1}}{y_i} = x_1$$

Note 1. If in an inept choice of solution it appears that $C_1=0$ and $C_2 \neq 0$, then the limit (6) will be equal to the next largest (in modulus) root of equation (1).

Note 2. If for the solution y_i the ratio $\frac{y_{i+1}}{y_i}$ oscillates without tending to a limit, then we may suspect that (1) has complex roots which are largest in modulus.

Note 3. Making the change of variable

$$x = \frac{1}{z}$$

in (1), it is possible, using Bernoulli's method, to find the least-modulus nonzero root of equation (1).

Thus, as an approximation to the maximum-modulus root x_1 of equation (1), we can use the formula

$$x_1 \approx \frac{y_i}{y_{i-1}}$$

where i is sufficiently large.

In a practical application of the Bernoulli method, one can specify arbitrary numbers $y_0, y_1, \ldots, y_{n-1}$ and then, using the formula

$$y_{n+i} = -\frac{1}{a_0}(a_n y_i + a_{n-1} y_{i-1} + \ldots + a_1 y_{n+i-1}) \qquad (i = 0, 1, 2, \ldots)$$

compute the sequence of numbers $y_n, y_{n+1}, y_{n+2}, \ldots$ and the ratios $\frac{y_n}{y_{n-1}}, \frac{y_{n+1}}{y_n}, \frac{y_{n+2}}{y_{n+1}}, \ldots$. If, as i increases, the ratio $\frac{y_{n+i}}{y_{n+i-1}}$ exhibits a tendency to approach some number ξ, the latter is taken as the maximum-modulus root x_1 of equation (1), otherwise it is extremely possible that the equation (1) has several maximum-modulus roots, or (less probable) that the coefficient $C_1 = 0$ for the initial sequence of numbers $y_0, y_1, \ldots, y_{n-1}$.

If a crude value α of the largest, in modulus, root x_1 is known, then it is advantageous, in order to accelerate the convergence of the process, to put

$$y_0 = 1, \quad y_1 = \alpha, \quad \ldots, \quad y_{n-1} = \alpha^{n-1}$$

Note that the Bernoulli method reduces to a repetitive sequence of operations of the same type and therefore is very suitable for machine computation.

The initial values y_i $(i=0,\ 1,\ \ldots,\ n-1)$ can, generally speaking, be taken in arbitrary fashion. One ordinarily takes $y_0=y_1=\ldots$ $\ldots=y_{n-2}=0;\ y_{n-1}=1$. Hildebrand [9] has suggested choosing y_i so that all the coefficients C_i in (4) are equal to unity. In that case, the process $\frac{y_i}{y_{i-1}}$ definitely converges as $i\to\infty$ provided there is a unique maximum-modulus root of equation (1).

The Bernoulli method can also be used to compute the complex roots of equation (1) [10].

Example. Find the maximum-modulus root x_1 of the equation

$$x^5+5x^4-5=0$$

Solution. The appropriate difference equation is of the form

$$y_{i+5}=5\,(y_i-y_{i+4})\qquad (i=0,\ 1,\ 2,\ \ldots)\tag{9}$$

In arbitrary fashion we take the values

$$y_0=0,\quad y_1=0,\quad y_2=0,\quad y_3=0,\quad y_4=1$$

By formula (9) we compute the values of y_i for $i\geqslant 5$. These values are listed in Table 11.

TABLE 11

FINDING THE ROOTS OF AN ALGEBRAIC EQUATION BY THE BERNOULLI METHOD

i	y_i	$\frac{y_i}{y_{i-1}}$	i	y_i	$\frac{y_i}{y_{i-1}}$
5	—5	—5	10	15,575	—4.992
6	25	—5	11	—77,750	—4.928
7	—125	—5	12	388,125	—4.99196
8	625	—5	13	—1,937,500	—4.991948
9	—3120	—4.992			

Terminating with y_{13}, we have

$$x_1\approx\frac{y_{13}}{y_{12}}=-\frac{1,937,500}{388,125}=-4.991948$$

whence, taking y_{12} into account, we can approximately put

$$x_1=-4.99195$$

In conclusion, it is worth noting that new methods of solving algebraic equations have recently appeared with convenient computational schemes (Lin's method, the method of N. V. Paluver, and others) [10].

REFERENCES FOR CHAPTER 5

[1] *A. G. Kurosh, Course of Higher Algebra* (translated from the Russian, Mir Publishers, 1972), Chapters 7 and 8.
[2] *G. M. Shapiro, Higher Algebra,* 1938, Chapters III and VI (in Russian).
[3] *D. Grave, Elements of Higher Algebra,* 1914, Chapter X (in Russian).
[4] *B. A. Fuks and B. V. Shabat, Functions of a Complex Variable,* 1949, Chapter VII (in Russian).
[5] *A. N. Krylov, Lectures on Approximate Computations,* 1933, Chapter II (in Russian).
[6] *J. B. Scarborough, Numerical Mathematical Analysis,* 1955, Chapter X.
[7] *B. K. Mlodzeevsky, Solution of Numerical Equations,* 1924, Chapter IV (in Russian).
[8] *A. O. Gelfond, Calculus of Finite Differences,* 1952, Chapter V (in Russian).
[9] *F. B. Hildebrand, Introduction to Numerical Analysis,* 1956.
[10] *V. L. Zaguskin, Handbook on Numerical Methods of Solving Algebraic and Transcendental Equations,* 1960 (in Russian).

Chapter 6

ACCELERATING THE CONVERGENCE OF SERIES

6.1 ACCELERATING THE CONVERGENCE OF NUMERICAL SERIES

We say that the series

$$a_1 + a_2 + \dots + a_n + \dots \tag{1}$$

converges slowly if we have to take a large number of terms of the series in order to obtain the sum to the required degree of accuracy. For instance, suppose we have to find the sum of the series

$$S = \frac{1}{1^2} + \frac{1}{2^2} + \dots + \frac{1}{n^2} + \dots \tag{2}$$

to within 10^{-6}. For the nth remainder of the series we have

$$R_n < \int\limits_n^{\infty} \frac{dx}{x^2} = \frac{1}{n}$$

Thus, our accuracy will be assured if we take the sum of 1,000,000 terms of the series, but this is impossible in any practical sense. Therefore, in solving this problem we regard the series (2) as a slowly convergent series.

To find the sum directly of a slowly convergent series to a specified accuracy ε is, generally speaking, an arduous task or practically impossible. Of importance, therefore, are transformations of the series which accelerate the convergence. We shall examine here the *Kummer transformation* [3], [4], which will be found to be useful in a number of cases.

Let the series (1) converge and let the sum be A. We choose an auxiliary convergent series

$$b_1 + b_2 + \dots + b_n + \dots \qquad (b_n \neq 0) \tag{2'}$$

the sum of which, B, is known, a series such that

$$\lim_{n \to \infty} \frac{a_n}{b_n} = q \neq 0 \tag{3}$$

Then we have the obvious equation

$$\sum_{n=1}^{\infty} a_n = q\sum_{n=1}^{\infty} b_n + \sum_{n=1}^{\infty} (a_n - qb_n)$$

or

$$A = qB + \sum_{n=1}^{\infty} (a_n - qb_n) \tag{4}$$

In particular, if $a_n \sim b_n$, then $q = 1$ and we have

$$A = B + \sum_{n=1}^{\infty} (a_n - b_n) \tag{4'}$$

Thus, finding the sum of the series (1) is replaced, in the general case, by finding the sum of the series

$$\sum_{n=1}^{\infty} (a_n - qb_n) \tag{5}$$

The remainder of the series (5), $\overline{R}_N$, may be written as

$$\overline{R}_N = \sum_{n=N+1}^{\infty} (a_n - qb_n) = \sum_{n=N+1}^{\infty} \left(1 - q\frac{b_n}{a_n}\right) a_n = \sum_{n=N+1}^{\infty} \varepsilon_n a_n$$

where $\varepsilon_n = 1 - q\frac{b_n}{a_n} \to 0$ as $n \to \infty$.

For this reason, in the general case, the series (5) converges faster than the original series (1). The main difficulty in applying the Kummer transformation consists in choosing a suitable auxiliary series (2′).

We demonstrate the application of this transformation for the positive series (1) whose terms a_n are rational functions of an integral variable n; that is,

$$a_n = \frac{\alpha_0 n^p + \alpha_1 n^{p-1} + \ldots + \alpha_p}{\beta_0 n^q + \beta_1 n^{q-1} + \ldots + \beta_q} \quad (n = 1, 2, \ldots) \tag{6}$$

where p and q are nonnegative integers and $\alpha_0 > 0$, $\beta_0 > 0$. For convergence of a series with general term (6), it is necessary and sufficient that the inequality $q \geqslant p + 2$ hold true.

In this case

$$a_n = O\left(\frac{1}{n^2}\right)^{1)}$$

(at least!).

---

[1] We say that a_n is an infinitesimal of order not less than m with respect to $\frac{1}{n}$:

Consider the auxiliary series

$$S^{(m)} = \sum_{n=1}^{\infty} \frac{1}{n(n+1)\dots(n+m)} \qquad (m=1,\ 2,\ \dots) \tag{7}$$

Since

$$\frac{1}{n(n+1)\dots(n+m)} = \frac{1}{m}\left[\frac{1}{n(n+1)\dots(n+m-1)} - \frac{1}{(n+1)(n+2)\dots(n+m)}\right]$$

it follows that

$$S_N^{(m)} = \sum_{n=1}^{N} \frac{1}{n(n+1)\dots(n+m)} = \frac{1}{m}\left[\frac{1}{1\cdot 2\dots m} - \frac{1}{(N+1)(N+2)\dots(N+m)}\right]$$

Hence

$$S^{(m)} = \lim_{N\to\infty} S_N^{(m)} = \frac{1}{m\,m!} \tag{8}$$

Utilizing Stirling's idea, represent the general term of the series as defined by formula (6) in the form of a finite sum of inverse factorials

$$a_n = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)} + a_n^{(m)}$$

where A_1, A_2, $\dots$, A_m are undetermined coefficients and $a_n^{(m)}$ is the remainder term. Select the coefficients A_i $(i=1,\ 2,\ \dots,\ m)$ so that

$$a_n^{(m)} = O\left(\frac{1}{n^{2+m}}\right)$$

$$a_n = O\left(\frac{1}{n^m}\right)$$

if

$$\lim_{n\to\infty} \frac{a_n}{\left(\frac{1}{n}\right)^m} = c \neq \infty$$

If $c \neq 0$, then a_n is an infinitesimal of order exactly m with respect to $\frac{1}{n}$.

For this purpose it suffices to determine the coefficients A_i successively from the formulas

$$\left.\begin{aligned} A_1 &= \lim_{n\to\infty} n(n+1)a_n, \\ A_2 &= \lim_{n\to\infty}\left[a_n-\frac{A_1}{n(n+1)}\right] n(n+1)(n+2), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ A_m &= \lim_{n\to\infty}\left[a_n-\sum_{i=1}^{m-1}\frac{A_i}{n(n+1)\dots(n+i)}\right] n(n+1)\dots(n+m) \end{aligned}\right\} \quad (9)$$

In accordance with the general scheme, we take for the auxiliary series (2)

$$B=\sum_{n=1}^{\infty} b_n=\sum_{n=1}^{\infty}\left[\frac{A_1}{n(n+1)}+\frac{A_2}{n(n+1)(n+2)}+\dots+\frac{A_m}{n(n+1)\dots(n+m)}\right]=$$

$$=A_1S^{(1)}+A_2S^{(2)}+\dots+A_mS^{(m)}=\frac{A_1}{1\cdot 1!}+\frac{A_2}{2\cdot 2!}+\dots+\frac{A_m}{m\cdot m!} \quad (10)$$

It is obvious that

$$\lim_{n\to\infty}\frac{a_n}{b_n}=1$$

and

$$S=\sum_{n=1}^{\infty} a_n=B+\sum_{n=1}^{\infty} a_n^{(m)} \quad (11)$$

Since the rapid convergence of the supplementary series $\sum_{n=1}^{\infty} a_n^{(m)}$ is, generally speaking, revealed only for a sufficiently large n, it is convenient in a practical sense to perform the indicated transformation beginning with some term a_{p+1} of the series. Assuming

$$S=\sum_{n=1}^{p} a_n+\sum_{n=p+1}^{\infty} a_n=S_p+\sum_{n=p+1}^{\infty} a_n$$

we have

$$S=S_p+\sum_{n=p+1}^{\infty}\left[\frac{A_1}{n(n+1)}+\frac{A_2}{n(n+1)(n+2)}+\dots+\frac{A_m}{n(n+1)\dots(n+m)}+a_n^{(m)}\right]=$$

$$=S_p+A_1\sum_{n=p+1}^{\infty}\left(\frac{1}{n}-\frac{1}{n+1}\right)+\frac{A_2}{2}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)}-\frac{1}{(n+1)(n+2)}\right]+\dots$$

$$\dots+\frac{A_m}{m}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)\dots(n+m-1)}-\frac{1}{(n+1)\dots(n+m)}\right]+\sum_{n=p+1}^{\infty} a_n^{(m)}=$$

$$= S_p + A_1 \cdot \frac{1}{p+1} + \frac{A_2}{2} \cdot \frac{1}{(p+1)(p+2)} + \dots$$
$$\dots + \frac{A_m}{m} \cdot \frac{1}{(p+1)\dots(p+m)} + \sum_{n=p+1}^{\infty} a_n^{(m)}$$

In particular, as $m \to \infty$, we obtain *Stirling's expansion* (taking into account that $a_n^{(m)} \rightrightarrows 0$):

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{p} a_n + A_1 \cdot \frac{1}{p+1} + \frac{A_2}{2} \cdot \frac{1}{(p+1)(p+2)} + \dots$$
$$\dots + \frac{A_m}{m} \cdot \frac{1}{(p+1)(p+2)\dots(p+m)} + \dots$$

Example. Find the sum of the series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2+1} \tag{12}$$

to within 0.001.

Solution. Setting

$$\frac{1}{n^2+1} = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + a_n^{(2)}$$

we have

$$A_1 = \lim_{n \to \infty} \frac{n(n+1)}{n^2+1} = 1,$$

$$A_2 = \lim_{n \to \infty} \left[\frac{1}{n^2+1} - \frac{1}{n(n+1)}\right] n(n+1)(n+2) = \lim_{n \to \infty} \frac{(n-1)(n+2)}{n^2+1} = 1$$

Hence

$$a_n^{(2)} = \frac{1}{n^2+1} - \frac{1}{n(n+1)} - \frac{1}{n(n+1)(n+2)} =$$
$$= \frac{n^3+3n^2+2n-n^3-2n^2-n-2-n^2-1}{n(n+1)(n+2)(n^2+1)} = \frac{n-3}{n(n+1)(n+2)(n^2+1)}$$

On the basis of formulas (10) and (11) we get

$$S = \frac{1}{1 \cdot 1!} + \frac{2}{2 \cdot 2!} + \sum_{n=1}^{\infty} \frac{n-3}{n(n+1)(n+2)(n^2+1)} \tag{12'}$$

Since for $n \geqslant 3$ we have

$$\frac{n-3}{n(n+1)(n+2)(n^2+1)} \leqslant \frac{1}{n^4}$$

it follows that

$$\rho_N = \sum_{n=N+1}^{\infty} \frac{n-3}{n(n+1)(n+2)(n^2+1)} < \int_N^{\infty} \frac{dx}{x^4} = \frac{1}{3N^3} < \frac{1}{2} \cdot 0.001$$

And from this it follows that the number of terms in the sum (12′) may be taken $N=10$; these summands must be computed to four decimal places in the narrow sense. We thus have

$$S \approx 1.25+(-0.1667)+(-0.0083)+0+0.0005+0.0004+ \\ +0.0002+0.0002+0.0001+0.0001+0.0001=1.0766$$

Noting that the sum of the first four summands is exact, for the absolute error of the result we get the estimate

$$\Delta < \frac{1}{3}\cdot 10^{-3}+7\cdot\frac{1}{2}\cdot 10^{-4} < 0.7\cdot 10^{-3}$$

Rounding, we find

$$S \approx 1.077$$

with the limiting absolute error

$$\bar{\Delta}=0.7\cdot 10^{-3}+0.4\cdot 10^{-3}=1.1\cdot 10^{-3}$$

Note that for the remainder of the given series (12) we have the estimate

$$R_N < \int_N^\infty \frac{dx}{x^2+1} < \int_N^\infty \frac{dx}{x^2}=\frac{1}{N} \leqslant \frac{1}{2}\cdot 0.001$$

whence $N \geqslant 2000$, which means that without the transformation we would need about 2000 terms of the series to attain the same accuracy.

Note. We could also use the following series for an approximate computation of the sum of the series (1) with general term (6):

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6},\quad \sum_{n=1}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{90},\quad \sum_{n=1}^{\infty}\frac{1}{n^6}=\frac{\pi^6}{945},\ \text{etc.}$$

Generally speaking,

$$\sum_{n=1}^{\infty}\frac{1}{n^{2p}}=\frac{(-1)^{p-1}}{2}\cdot\frac{B_{2p}(2\pi)^{2p}}{(2p)!}$$

where B_n $(n=1, 2, \ldots)$ are Bernoulli numbers [5], [6] defined by the symbolic formula

$$(B+1)^n-B^n=0$$

in which, after expanding by the binomial theorem, we put $B^n=B_n$. In particular, we have

$$B_2=\frac{1}{6},\quad B_4=-\frac{1}{30},\quad B_6=\frac{1}{42},\quad B_8=-\frac{1}{30},\quad B_{10}=\frac{5}{66}$$

(see Sec. 16.11).

6.2 ACCELERATING THE CONVERGENCE OF POWER SERIES BY THE EULER-ABEL METHOD

Consider the convergent power series

$$f(x)=\sum_{n=0}^{\infty} a_n x^n \tag{1}$$

where $f(x)$ is the sum of the series.

Let the radius of convergence R of series (1) be finite and nonzero. Without loss of generality, we can take it that $R=1$.[1]

Write the series (1) as

$$f(x)=a_0+x\varphi(x) \tag{2}$$

where

$$\varphi(x)=\sum_{n=1}^{\infty} a_n x^{n-1}=\sum_{n=0}^{\infty} a_{n+1}x^n \tag{3}$$

Multiplying both sides of (3) by the binomial $1-x$, we get

$$(1-x)\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n-\sum_{n=0}^{\infty} a_{n+1}x^{n+1} \tag{4}$$

Assuming $n+1=m$ in the second sum and noting that the sum does not depend on the summation index, we have

$$\sum_{n=0}^{\infty} a_{n+1}x^{n+1}=\sum_{m=1}^{\infty} a_m x^m=\sum_{n=1}^{\infty} a_n x^n$$

Therefore

$$(1-x)\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n-\sum_{n=1}^{\infty} a_n x^n=$$

$$=a_0+\sum_{n=0}^{\infty}(a_{n+1}-a_n)x^n=a_0+\sum_{n=0}^{\infty}\Delta a_n x^n$$

where

$$\Delta a_n=a_{n+1}-a_n \qquad (n=0,\ 1,\ 2,\ \ldots)$$

are *finite differences of the first order* of the coefficients a_n (for more on finite differences see Sec. 14.1). Thus from formulas (3) and (4) we derive

$$\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n=\frac{a_0}{1-x}+\frac{1}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n$$

---

[1] Indeed, if $0<R<\infty$ and $R\neq 1$, then, assuming $t=\frac{x}{R}$, we get a power series in the variable t with radius of convergence $\rho=1$.

and, hence,

$$f(x) = a_0 + \frac{a_0 x}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty} \Delta a_n x^n = \frac{a_0}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty} \Delta a_n x^n$$

that is

$$\sum_{n=0}^{\infty} a_n x^n = \frac{a_0}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty} \Delta a_n x^n \tag{5}$$

This transformation of a power series is called the *Euler-Abel transformation*. Analogously, applying the Euler-Abel transformation to the power series $\sum\limits_{n=0}^{\infty} \Delta a_n x^n$, we find

$$\sum_{n=0}^{\infty} \Delta a_n x^n = \frac{\Delta a_0}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty} \Delta^2 a_n x^n$$

where

$$\Delta^2 a_n = \Delta(\Delta a_n) = \Delta a_{n+1} - \Delta a_n$$

are *finite differences of the second order* of the coefficients a_n, whence, on the basis of formula (5), we get

$$\sum_{n=0}^{\infty} a_n x^n = \frac{a_0}{1-x} + \frac{x}{1-x}\left(\frac{\Delta a_0}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty} \Delta^2 a_n x^n\right) =$$

$$= \frac{a_0}{1-x} + \frac{x\Delta a_0}{(1-x)^2} + \left(\frac{x}{1-x}\right)^2 \sum_{n=0}^{\infty} \Delta^2 a_n x^n$$

Repeating the Euler-Abel transformation p times in succession, we obtain

$$\sum_{n=0}^{\infty} a_n x^n = \frac{a_0}{1-x} + \frac{x\Delta a_0}{(1-x)^2} + \ldots + \frac{x^{p-1}\Delta^{p-1} a_0}{(1-x)^p} + \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty} \Delta^p a_n x^n$$

where

$$\Delta^p a_n = \Delta^{p-1} a_{n+1} - \Delta^{p-1} a_n \qquad (n = 0, 1, 2, \ldots)$$

are *finite differences of the* p*th order* of the coefficients a_n, and $\Delta^k a_0$ ($k = 0, 1, 2, \ldots$) are the successive finite differences of the coefficients a_n for $n = 0$. Thus

$$f(x) = \sum_{k=0}^{p-1} \Delta^k a_0 \frac{x^k}{(1-x)^{k+1}} + \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty} \Delta^p a_n x^n \tag{6}$$

where $\Delta^0 a_0 = a_0$. Formula (6) is advantageously used when the finite differences $\Delta^p a_n$ are of higher order of decay, as $n \to \infty$, than the coefficients a_n. This occurs frequently. For example, if $a_n=\frac{1}{n}$, then we get

$$\Delta a_n=\frac{1}{n+1}-\frac{1}{n}=-\frac{1}{n(n+1)}$$

That is, Δa_n decreases faster than a_n as $n\to\infty$.

In particular, if $a_n=P(n)$, where $P(n)$ is an integral polynomial of degree $p-1$, then formula (6) yields the sum of the series

$$\sum_{n=0}^{\infty} P(n)\,x^n=\sum_{k=0}^{p-1}\Delta^k P(0)\frac{x^k}{(1-x)^{k+1}} \qquad (|x|<1) \tag{7}$$

in closed form, since $\Delta^p P(n)=0$.

Formula (6) becomes meaningless when $x=1$. The Euler-Abel transformation may be modified to accommodate this case. Setting $x=-t$, we have

$$f(x)=\sum_{n=0}^{\infty} a_n(-t)^n=\sum_{n=0}^{\infty}(-1)^n a_n t^n=$$
$$=\sum_{k=0}^{p-1}\Delta^k[(-1)^n a_n]_{n=0}\frac{t^k}{(1-t)^{k+1}}+\left(\frac{t}{1-t}\right)^p\sum_{n=0}^{\infty}\Delta^p[(-1)^n a_n]\,t^n$$

Returning to the earlier variable, we obtain

$$f(x)=\sum_{k=0}^{p-1}(-1)^k\Delta^k[(-1)^n a_n]_{n=0}\frac{x^k}{(1+x)^k}+$$
$$+\left(\frac{x}{1+x}\right)^p\sum_{n=0}^{\infty}(-1)^{n+p}\Delta^p[(-1)^n a_n]\,x^n \tag{8}$$

Formula (8) is meaningful for $x=1$ as well.

Example 1. Find the sum of the series

$$f(x)=\sum_{n=0}^{\infty}\frac{x^n}{(n+1)(n+2)} \tag{9}$$

to within 0.001 when $x=-1$.

Solution. Apply the Euler transformation twice ($p=2$) to get

$$a_n=\frac{1}{(n+1)(n+2)},$$

$$\Delta a_n=a_{n+1}-a_n=\frac{1}{(n+2)(n+3)}-\frac{1}{(n+1)(n+2)}=$$

$$=-\frac{2}{(n+1)(n+2)(n+3)},$$

$$\Delta^2 a_n=\Delta a_{n+1}-\Delta a_n=-\frac{2}{(n+2)(n+3)(n+4)}+$$

$$+\frac{2}{(n+1)(n+2)(n+3)}=\frac{6}{(n+1)(n+2)(n+3)(n+4)}$$

Hence

$$a_0=\frac{1}{1\cdot 2},\qquad \Delta a_0=-\frac{2}{1\cdot 2\cdot 3}$$

From this, on the basis of formula (6), we get

$$f(-1)=\frac{1}{1\cdot 2}\cdot\frac{1}{2}+\frac{2}{1\cdot 2\cdot 3}\cdot\frac{1}{4}+$$

$$+\left(-\frac{1}{2}\right)^2\sum_{n=0}^{\infty}\frac{6}{(n+1)(n+2)(n+3)(n+4)}(-1)^n=$$

$$=\frac{1}{4}+\frac{1}{12}+\frac{3}{2}\cdot\frac{1}{24}-\frac{3}{2}\cdot\frac{1}{120}+\frac{3}{2}\cdot\frac{1}{360}-\frac{3}{2}\cdot\frac{1}{840}+$$

$$+\frac{3}{2}\cdot\frac{1}{1680}-\frac{3}{2}\cdot\frac{1}{3024}+\frac{3}{2}\cdot\frac{1}{5040}-\dots \qquad (10)$$

The series (10) is an alternating series with terms decreasing monotonically in modulus. Therefore, if we stop with the term

$$\frac{3}{2}\cdot\frac{1}{3024}=\frac{1}{2016}$$

then the remainder R of the series will not exceed (in modulus) the first discarded term:

$$|R|<\frac{3}{2}\cdot\frac{1}{5040}=\frac{1}{3360}<3\cdot 10^{-4}$$

Thus, taking two extra digits, we have

$$f(-1)=0.25000+0.08333+0.06250-0.01250+$$

$$+0.00417-0.00179+0.00089-0.00050=0.38610$$

with absolute error

$$\Delta<5\cdot\frac{1}{2}\cdot 10^{-5}+3\cdot 10^{-4}<4\cdot 10^{-4}$$

Rounding this number to three decimals, we get the approximate value $f(-1)=0.386$ with limiting absolute error

$$\Delta < 4\cdot 10^{-4}+1\cdot 10^{-4}=\frac{1}{2}\cdot 10^{-3}$$

The exact value of the sum is

$$f(-1)=2\ln 2-1=0.38630\ldots$$

It will be noted that if one computed the number $f(-1)$ directly, using the series (9), roughly forty-five terms of the series would be needed to attain the required accuracy.

Example 2. Find the sum of the series

$$S(x)=\sum_{n=0}^{\infty}(n^2+n+1)x^n$$

Solution. We have

$$P(n)=n^2+n+1$$

Construct Table 12.

TABLE 12

TABLE OF FINITE DIFFERENCES

n	$P(n)$	$\Delta P(n)$	$\Delta^2 P(n)$
0	1	2	2
1	3	4	
2	7		

Formula (7) yields

$$S(x)=\frac{1}{1-x}+\frac{2x}{(1-x)^2}+\frac{2x^2}{(1-x)^3}$$

for $|x|<1$.

6.3 ESTIMATES OF FOURIER COEFFICIENTS

The Fourier trigonometric series of a given function $f(x)\ (-\pi<x<\pi)$[1] is the series

$$\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cos nx+b_n\sin nx) \tag{1}$$

---

[1] For the sake of simplicity of formulation we consider the function defined on the interval $[-\pi, \pi]$. The general case of the function $\varphi(t)$ defined on the interval $[a, b]$ may be reduced to ours by means of the linear substitution

$$t=\frac{b+a}{2}+\frac{b-a}{2\pi}x$$

the coefficients a_n, b_n of which [*Fourier coefficients* of the function $f(x)$] are computed from the formulas

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \qquad (n = 0,\ 1,\ \ldots), \tag{2}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \qquad (n = 1,\ 2,\ \ldots) \tag{2'}$$

A sufficient condition for the existence of a Fourier series of a function $f(x)$ is the integrability of the function on the interval $[-\pi, \pi]$. In this case, the Fourier coefficients (2) and (2′) have definite finite values.

It may happen that the resulting Fourier series diverges or converges to a different function. We give without proof [1], [7] the conditions under which a Fourier trigonometric series converges to a function $f(x)$ at all points of continuity of the function.

Convergence theorem. *If a function* $f(x)$ *is piecewise continuous and piecewise differentiable on the interval* $[-\pi, \pi]$, *then its Fourier series converges on the whole number axis and its sum* $S(x)$ *is a periodic function, with period* 2π, *equal to*

$$S(x_0) = \frac{f(x_0-0)+f(x_0+0)}{2} \tag{3}$$

at any point $x_0 \in (-\pi, \pi)$ *and* $S(\pm\pi) = 2^{-1}[f(-\pi+0)+f(\pi-0)]$.

In particular, $S(x_0) = f(x_0)$ if the function is continuous at the point $x = x_0$, that is, if $f(x_0-0) = f(x_0+0) = f(x_0)$.

If, besides, the function $f(x)$ is periodic with period 2π, then its Fourier series converges for every value x_0 and has the sum (3).

If the conditions of the convergence theorem are fulfilled, then it is obvious that $a_n \to 0$ and $b_n \to 0$ as $n \to \infty$. We give more exact estimates of the Fourier coefficients by imposing certain restrictions on the behaviour of the function $f(x)$.

Definition. *We say that a function* $f(x)$ *specified on the interval* $[-\pi, \pi]$ *belongs to periodicity class* $\tilde{C}^{(m)}$ *if*:

(1) $f(x)$ *is continuous on* $[-\pi, \pi]$ *together with its derivatives up to order* m *inclusive*;

(2) $f^{(k)}(-\pi+0) = f^{(k)}(\pi-0)$ *for* $k = 0, 1, 2, \ldots, m$, *that is, at the endpoints of* $[-\pi, \pi]$ *the values of the function* $f(x)$ *must coincide with its first* m *derivatives.*

From Conditions (1) and (2) it follows that a periodic continuation of $f(x)$ belongs to the class $C^{(m)}(-\infty, +\infty)$.

Lemma. *If a function* $f(x)$ *belongs to periodicity class* $\tilde{C}^{(m)}$ *on the interval* $[-\pi, \pi]$ *(more briefly,* $f(x) \in \tilde{C}^{(m)}[-\pi, \pi]$*), then its Fourier coefficients* a_n *and* b_n *are infinitesimals, as* $n \to \infty$, *of order* *higher than* m *with respect to* $\frac{1}{n}$, *that is,*

$$a_n = o\left(\frac{1}{n^m}\right), \quad b_n = o\left(\frac{1}{n^m}\right)^{1)}$$

Proof. Integrate the right members of the following equations by parts m times:

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \qquad (n=0,\ 1,\ \ldots), \tag{4}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \qquad (n=1,\ 2,\ \ldots) \tag{4'}$$

Putting $u=f(x)$ and $dv=\cos nx\,dx$, we find $du=f'(x)\,dx$ and $v=\frac{1}{n}\sin nx$. Thus, by the formula for integration by parts, we have

$$a_n = \frac{1}{\pi}\left[\frac{1}{n}f(x)\sin nx\right]_{-\pi}^{\pi} - \frac{1}{\pi n}\int_{-\pi}^{\pi} f'(x)\sin nx\,dx =$$

$$= \frac{1}{\pi n}\int_{-\pi}^{\pi} f'(x)\cos\left(\frac{\pi}{2}+nx\right)dx$$

Applying integration by parts once again and noting that $f'(-\pi)=f'(\pi)$, we get

$$a_n = \frac{1}{\pi n}\left\{\left[\frac{1}{n}f'(x)\sin\left(\frac{\pi}{2}+nx\right)\Big|_{-\pi}^{\pi} + \frac{1}{n}\int_{-\pi}^{\pi}(f''x)\cos\left(\frac{\pi}{2}\cdot 2+nx\right)dx\right]\right\} =$$

$$= \frac{1}{\pi n^2}\int_{-\pi}^{\pi} f''(x)\cos\left(\frac{\pi}{2}\cdot 2+nx\right)dx$$

and so forth.

After an m-fold integration by parts we have, in formulas (4) and (4′),

$$a_n = \frac{1}{\pi n^m}\int_{-\pi}^{\pi} f^{(m)}(x)\cos\left(\frac{\pi}{2}\cdot m+nx\right)dx$$

---

1) The notation $a_n = o\left(\frac{1}{n^m}\right)$ means that $\lim\limits_{n\to\infty}\left(\frac{a_n}{\frac{1}{n^m}}\right)=0$.

Similarly

$$b_n = \frac{1}{\pi n^m} \int_{-\pi}^{\pi} f^{(m)}(x) \sin\left(\frac{\pi}{2} \cdot m + nx\right) dx$$

The integrals

$$\varepsilon_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f^{(m)}(x) \cos\left(\frac{\pi}{2} \cdot m + nx\right) dx$$

and

$$\varepsilon_n' = \frac{1}{\pi} \int_{-\pi}^{\pi} f^{(m)}(x) \sin\left(\frac{\pi}{2} \cdot m + nx\right) dx$$

are, to within sign, the Fourier coefficients of the continuous (by hypothesis) function $f^{(m)}(x)$. As is well known, the Fourier coefficients of a continuous function tend to zero as their numbers increase without bound, irrespective of whether its Fourier series converges or not.[1] Therefore

$$\varepsilon_n \longrightarrow 0 \quad \text{and} \quad \varepsilon_n' \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty$$

But since

$$a_n = \frac{\varepsilon_n}{n^m} \quad \text{and} \quad b_n = \frac{\varepsilon_n'}{n^m}$$

it follows that the Fourier coefficients a_n and b_n of the function $f(x)$ are infinitesimals of higher order than $\frac{1}{n^m}$:

$$a_n = o\left(\frac{1}{n^m}\right), \qquad b_n = o\left(\frac{1}{n^m}\right)$$

A. N. Krylov used this result as the basis for a method of accelerating the convergence of the Fourier series.

Note. If $f^{(m)}(x)$ satisfies the conditions of the convergence theorem, then it is easy to prove that

$$\varepsilon_n = O\left(\frac{1}{n}\right) \quad \text{and} \quad \varepsilon_n' = O\left(\frac{1}{n}\right)$$

---

[1] This follows from the fact that for any piecewise continuous function $f(x)$ with Fourier coefficients a_n and b_n $(n=0, 1, 2, \ldots)$ the *Bessel inequality* [7] $\frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \leqslant \frac{1}{\pi} \int_{-\pi}^{\pi} f^2(x)\,dx$ is valid. Hence the series $\sum_{n=1}^{\infty} (a_n^2 + b_n^2)$ converges and $a_n \longrightarrow 0$, $b_n \longrightarrow 0$ as $n \longrightarrow \infty$;

In this case, a better estimate is obtained for the Fourier coefficients of the function $f(x)$:

$$a_n = O\left(\frac{1}{n^{m+1}}\right) \quad \text{and} \quad b_n = O\left(\frac{1}{n^{m+1}}\right)$$

6.4 ACCELERATING THE CONVERGENCE OF FOURIER TRIGONOMETRIC SERIES BY THE METHOD OF A. N. KRYLOV

Suppose a function $f(x)$ is piecewise continuous and has piecewise continuous derivatives $f^{(i)}(x)$ $(i=1, 2, \ldots, m)$ up to the mth order inclusive on the interval $[-\pi, \pi]$. Then by virtue of the convergence theorem of Sec. 6.3 the function $f(x)$ can be represented as a Fourier trigonometric series at all its points of continuity:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \tag{1}$$

where a_n and b_n are the Fourier coefficients defined by formulas (2) and (2′) of Sec. 6.3. In the general case, the coefficients a_n and b_n of the series (1) slowly approach zero; it is hard to use this series in any practical sense, all the more so is it inadmissible to differentiate the series (1) term by term, yet this is required in the solution of certain problems, in particular those involving the Fourier method.

The underlying idea of Krylov's method [8] is that one takes out of the function $f(x)$ an elementary function $g(x)$ (ordinarily a piecewise polynomial function) having the same discontinuities as $f(x)$; its derivatives $g^{(i)}(x)$ $(i=1, 2, \ldots, m)$ up to the mth order inclusive have the very same discontinuities as the corresponding derivatives $f^{(i)}(x)$ of the given function $f(x)$ and, what is more, $g(x)$ is such that

$$f^{(i)}(-\pi+0) - g^{(i)}(-\pi+0) = f^{(i)}(\pi-0) - g^{(i)}(\pi-0)$$
$$(i=0, 1, 2, \ldots, m)$$

In that case the difference

$$\varphi(x) = f(x) - g(x)$$

will belong to periodicity class $\tilde{C}^{(m)}$.

Denoting the Fourier coefficients of the function $\varphi(x)$ by α_n and β_n $(n=0, 1, 2, \ldots)$, we get

$$f(x) = g(x) + \left[\frac{\alpha_0}{2} + \sum_{n=1}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx)\right] \tag{2}$$

where α_n and β_n are infinitesimals, as $n \to \infty$, of order higher than m with respect to $\frac{1}{n}$, that is, the series (2) will, generally speaking, be a rapidly convergent series. This series can be differentiated term by term at least $m-2$ times.

We give a practical demonstration of how the auxiliary function $g(x)$ is constructed from the given function $f(x)$ [9]. To do this, on the interval $[-2\pi, 2\pi]$ we construct recursively a sequence of functions $\sigma_0(x)$, $\sigma_1(x)$, ..., $\sigma_m(x)$ having the property

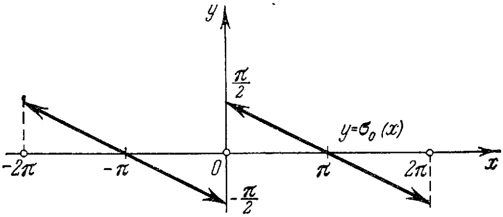
$$\sigma_k^{(k)}(+0)-\sigma_k^{(k)}(-0)=\pi \tag{3}$$

($k=0, 1, 2, \ldots, m$) and such that the derivatives $\sigma_k^{(j)}(x)$ ($j=0, 1, \ldots, k-1$) are continuous on $[-2\pi, 2\pi]$.

We define the function $\sigma_0(x)$ in the following manner:

$$\sigma_0(x)=\begin{cases} \frac{-\pi-x}{2} & \text{for} \quad -2\pi<x<0, \\ \frac{\pi-x}{2} & \text{for} \quad 0<x<2\pi, \\ 0 & \text{for} \quad x=-2\pi,\ 0,\ 2\pi \end{cases} \tag{4}$$

Fig. 46



Its graph is shown in Fig. 46. This function is odd and so its Fourier series contains only the sines of multiple arcs:

$$\sigma_0(x)=\sum_{n=1}^{\infty} b_n \sin nx$$

where

$$b_n=\frac{2}{\pi}\int_0^{\pi}\frac{\pi-x}{2}\sin nx\,dx=\frac{2}{\pi}\left(-\frac{\pi-x}{2}\cdot\frac{\cos nx}{n}\Big|_0^{\pi}-\frac{1}{2n}\int_0^{\pi}\cos nx\,dx\right)=$$

$$=\frac{2}{\pi}\left(\frac{\pi}{2n}-\frac{1}{2n^2}\sin nx\Big|_0^{\pi}\right)=\frac{1}{n}$$

Hence

$$\sigma_0(x)=\frac{\sin x}{1}+\frac{\sin 2x}{2}+\ldots+\frac{\sin nx}{n}+\ldots \tag{5}$$

It is obvious that the function $\sigma_0(x)$ has a discontinuity at the point $x=0$ with a jump equal to π:

$$\sigma_0(+0)-\sigma_0(-0)=\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)=\pi$$

and so the function

$$\psi(x)=\sigma_0(x-x_0) \quad (-\pi\leqslant x\leqslant\pi, \quad -\pi\leqslant x_0\leqslant\pi)$$

has the same jump at the point x_0 as the function $\sigma_0(x)$:

$$\psi(x_0+0)-\psi(x_0-0)=\pi$$

The point of discontinuity is the only one on the interval $[-\pi, \pi]$.

We define the function $\sigma_1(x)$ by the formula

$$\sigma_1(x)=c_1+\int_0^x \sigma_0(x)\,dx \tag{6}$$

where c_1 is a constant.

Integrating the series (5) termwise, we obtain

$$\sigma_1(x)=c_1+\int_0^x \sum_{n=1}^{\infty}\frac{\sin nx}{n}\,dx=c_1+\sum_{n=1}^{\infty}\frac{1}{n^2}-\sum_{n=1}^{\infty}\frac{\cos nx}{n^2} \tag{7}$$

We choose c_1 so that the constant term of series (7) is zero:

$$c_1+\sum_{n=1}^{\infty}\frac{1}{n^2}=0$$

whence

$$c_1=-\sum_{n=1}^{\infty}\frac{1}{n^2}$$

The series $\sum_{n=1}^{\infty}\frac{1}{n^2}$ is clearly the constant term of the Fourier series of the function $\int_0^x \sigma_0(x)\,dx$. From this, using the formula (4), we have

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{1}{\pi}\int_0^{\pi}dx\int_0^x \sigma_0(x)\,dx=\frac{1}{\pi}\int_0^{\pi}\left[\frac{\pi^2}{4}-\frac{(\pi-x)^2}{4}\right]dx=$$

$$=\frac{1}{\pi}\left(\frac{\pi^3}{4}-\frac{\pi^3}{12}\right)=\frac{\pi^2}{6}$$

Therefore

$$c_1=-\frac{\pi^2}{6}$$

Hence

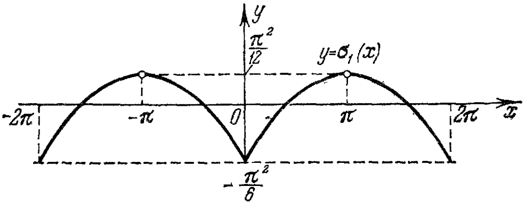
$$\sigma_1(x) = -\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} \tag{8}$$

and

$$\sigma_1(x) = \begin{cases} \displaystyle\int_0^x \frac{\pi - x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi - x)^2}{4} & \text{for} \quad 0 \leqslant x \leqslant 2\pi, \\ \displaystyle -\int_0^x \frac{\pi + x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi + x)^2}{4} & \text{for} \quad -2\pi \leqslant x \leqslant 0 \end{cases}$$

The graph of the function $\sigma_1(x)$ is shown in Fig. 47. The function $\sigma_1(x)$ is continuous on the interval $[-2\pi, 2\pi]$ but its derivative, $\sigma_1'(x) = \sigma_0(x)$, has a discontinuity at the point $x = 0$, and

Fig. 47



$$\sigma_1'(+0) - \sigma_0'(-0) = \pi$$

The following functions are defined in the same way:

$$\sigma_2(x) = \int_0^x \sigma_1(x)\,dx + c_2,$$

$$\sigma_3(x) = \int_0^x \sigma_2(x)\,dx + c_3,$$

. . . .

$$\sigma_m(x) = \int_0^x \sigma_{m-1}(x)\,dx + c_m$$

where the arbitrary constants $c_1, c_2, \ldots, c_m$ are chosen so that the constant term of the corresponding Fourier series is zero; that is, the constants c_k $(k = 1, 2, \ldots, m)$ are successively found from the conditions

$$\int_0^{\pi} \left[c_k + \int_0^x \sigma_{k-1}(x)\,dx \right] dx = 0$$

The functions $\sigma_k(x)$ $(k=1, 2, \ldots, m)$ and all the derivatives up to $(k-1)$th order inclusive are continuous on the interval $[-2\pi, 2\pi]$. Here, since $\sigma_k^{(k)}(x)=\sigma_0(x)$, it follows that

$$\sigma_k^{(k)}(+0)-\sigma_k^{(k)}(-0)=\pi \qquad (k=1, 2, \ldots, m)$$

that is, the derivative of kth order of the function $\sigma_k(x)$ has a discontinuity at $x=0$ with jump π. Whence it follows that the function $\psi_k(x)=\sigma_k(x-x_0)$ $(-\pi\leqslant x\leqslant\pi)$ obtained by shifting the function $\sigma_k(x)$ has a discontinuity of only the kth derivative at the point $x=x_0$:

$$\psi_k^{(k)}(x_0+0)-\psi_k^{(k)}(x_0-0)=\pi$$

Now let

$x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ be points of discontinuity of $f(x)$,
$x_1^{(1)}, x_2^{(1)}, \ldots, x_{k_1}^{(1)}$ be points of discontinuity of $f'(x)$,
. . . .
$x_1^{(m)}, x_2^{(m)}, \ldots, x_{k_m}^{(m)}$ be points of discontinuity of $f^{(m)}(x)$

Note that some of these points may be repeated.

We introduce the following notations for the corresponding jumps of the function and its derivatives:

$$f^{(l)}(x_j^{(l)}+0)-f^{(l)}(x_j^{(l)}-0)=h_j^{(l)}$$
$$(l=0, 1, \ldots, m; \quad j=1, 2, \ldots, k_l)$$

We define the function $g(x)$ (*jump function*) by the formula

$$g(x)=\sum_{s=1}^{s=k_0}\frac{h_s^{(0)}}{\pi}\sigma_0(x-x_s^{(0)})+\sum_{s=1}^{s=k_1}\frac{h_s^{(1)}}{\pi}\sigma_1(x-x_s^{(1)})+$$
$$+\ldots+\sum_{s=1}^{s=k_m}\frac{h_s^{(m)}}{\pi}\sigma_m(x-x_s^{(m)}) \qquad (9)$$

The function $g(x)$ has the following properties:

(1) at the points $x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ it has discontinuities, and the jumps at these points are equal to the jumps of the function $f(x)$ at the corresponding points:

$$g(x_j^{(0)}+0)-g(x_j^{(0)}-0)=\frac{h_j^{(0)}}{\pi}[\sigma_0(x_j-x_j+0)-$$
$$-\sigma_0(x_j-x_j-0)]=\frac{h_j^{(0)}}{\pi}\pi=h_j^{(0)}$$

(2) the derivative $g^{(l)}(x)$ $(l=1, 2, \ldots, m)$ is discontinuous at the points $x_1^{(l)}, x_2^{(l)}, \ldots, x_{k_l}^{(l)}$; also

$$g^{(l)}(x_j^{(l)}+0)-g^{(l)}(x_j^{(l)}-0)=\frac{h_j^{(l)}}{\pi}[\sigma_l(x_j^{(l)}-x_j^{(l)}+0)-$$
$$-\sigma_l(x_j^{(l)}-x_j^{(l)}-0)]=\frac{h_j^{(l)}}{\pi}\pi=h_j^{(l)}$$

that is,

$$g^{(l)}(x_j+0)-g^{(l)}(x_j-0)=f^{(l)}(x_j+0)-f^{(l)}(x_j-0)$$

(3) for $x\neq x_j^{(l)}$ the function $g(x)$ has continuous derivatives of all orders.

Suppose

$$\varphi(x)=f(x)-g(x) \tag{10}$$

By virtue of the first and second properties it follows that

$$\varphi^{(l)}(x_j^{(l)}+0)-\varphi^{(l)}(x_j^{(l)}-0)=0 \qquad (l=0, 1, 2, \ldots, m)$$

that is,

$$\varphi(x)\in\tilde{C}^{(m)}[-\pi, \pi]$$

Thus, the rapidly convergent Fourier series (2) can be used to expand the function $f(x)$. Note that by using the expansions

$$\sigma_0(x-x_s^{(0)})=\sum_{n=1}^{\infty}\frac{\sin n(x-x_s^{(0)})}{n},$$

$$\sigma_1(x-x_s^{(1)})=-\sum_{n-1}^{\infty}\frac{\cos n(x-x_s^{(1)})}{n^2},$$

$$\sigma_2(x-x_s^{(2)})=-\sum_{n=1}^{\infty}\frac{\sin n(x-x_s^{(2)})}{n^3},$$

. . . .

it is easy to write the Fourier expansion of the function $g(x)$. What we obtain finally is that the Fourier series of the function $f(x)$ consists of: (a) a slowly convergent part which is summable in elementary terms to the function $g(x)$, and (b) a rapidly convergent remainder which is the Fourier series of the function $\varphi(x)\in\tilde{C}^{(m)}[-\pi, \pi]$.

Note. If the limiting values of the function $f(x)$ or of its derivatives $f'(x), \ldots, f^{(k)}(x)$ $(k\leqslant m)$ do not coincide at the endpoints of the interval $[-\pi, \pi]$, i.e.,

$$f^{(l)}(-\pi+0)\neq f^{(l)}(\pi-0) \qquad (l=0, 1, 2, \ldots, k)$$

then the points $x=-\pi$ and $x=\pi$ are to be regarded as points of discontinuity of $f(x)$ or, respectively, of the derivatives $f^{(l)}(x)$.

Assuming that $f(x)$ is periodically continued beyond the limits of the interval $[-\pi, \pi]$ with period 2π, we find that the jump of the derivatives at the points $x=-\pi$ and $x=\pi$ is one and the same and is equal to

$$h^{(l)}=f^{(l)}(-\pi+0)-f^{(l)}(\pi-0)$$

By the periodicity of the function $\sigma_l(x)$ we have

$$\sigma_l(x+\pi)=\sigma_l(x-\pi)$$

On the interval $[-\pi, \pi]$, the function $\sigma_l^{(l)}(x+\pi)$ admits two discontinuity points ($x=-\pi$ and $x=\pi$) with one and the same jump π. And so we have to include only one endpoint, say $x=-\pi$; in formula (9). Indeed, by (9), the jump in the derivative $g^{(l)}(x)$ at the point $x=-\pi$ is equal to

$$g^{(l)}(-\pi+0)-g^{(l)}(-\pi-0)=\frac{h^{(l)}}{\pi}\left[\sigma^{(l)}(+0)-\sigma^{(l)}(-0)\right]=h^{(l)}$$

By the periodicity of $g^{(l)}(x)$, this derivative has the same jump at $x=\pi$ as well. Consequently, when forming the difference

$$f(x)-g(x)=\varphi(x)$$

where only the point $x=-\pi$ is taken into account, the discontinuity of the lth derivative of the function $\varphi(x)$ is removed both at the point $x=-\pi$ and at $x=\pi$.

Example. Using Krylov's method, accelerate the convergence of the Fourier series of the function (Fig. 48a)

$$f(x)=\begin{cases} x^2+1 & \text{for} \quad -\pi<x<0, \\ x^2 & \text{for} \quad 0<x<\pi \end{cases}$$

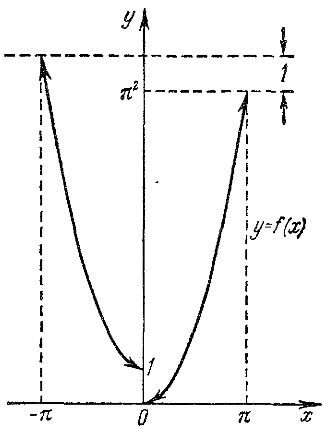


Fig. 48a

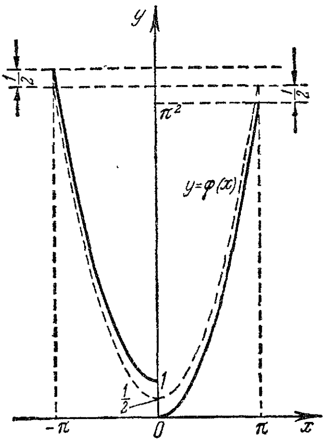


Fig. 48b

Solution. By virtue of the note, the function $f(x)$ has points of discontinuity on the interval $[-\pi, \pi]$: $x_1=-\pi$, $x_2=0$, $x_3=\pi$. Computing the Fourier coefficients, we obtain

$$a_0=1+\frac{2\pi^2}{3}, \quad a_n=\frac{4}{n^2}(-1)^n, \quad b_n=\begin{cases} -\frac{2}{\pi n} & \text{for } n \text{ odd,} \\ 0 & \text{for } n \text{ even} \end{cases}$$

Hence, the Fourier series of the function $f(x)$ has the form

$$f(x)=\frac{1}{2}+\frac{\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos nx-\frac{2}{\pi}\sum_{k=0}^{\infty}\frac{1}{2k+1}\sin(2k+1)x \qquad (11)$$

The convergence of (11) is poor since the coefficients $b_n=O\left(\frac{1}{n}\right)$ decrease slowly. From the function $f(x)$ we isolate the jump function $g(x)$ so that $\varphi(x)=[f(x)-g(x)]\in\tilde{C}^{(m)}\ [-\pi, \pi]$.

Let us compute the jumps $h_j^{(0)}$ of zero kind at the points $x_j\,(j=1, 2, 3)$:

$$\begin{aligned} h_1^{(0)}&=f(-\pi+0)-f(\pi-0)=(\pi^2+1)-\pi^2=1, \\ h_2^{(0)}&=f(+0)-f(-0)=0-1=-1, \\ h_3^{(0)}&=h_1^{(0)}=1 \end{aligned}$$

On the basis of formula (9) and taking into account the note, we get

$$g(x)=\frac{1}{\pi}\cdot\sigma_0(x+\pi)-\frac{1}{\pi}\cdot\sigma_0(x)$$

or

$$g(x)=\frac{1}{\pi}\cdot\frac{\pi-(x+\pi)}{2}+\frac{1}{\pi}\cdot\frac{\pi+x}{2}=\frac{1}{2}$$

for $-\pi<x<0$ and

$$g(x)=\frac{1}{\pi}\cdot\frac{\pi-(x+\pi)}{2}-\frac{1}{\pi}\cdot\frac{\pi-x}{2}=-\frac{1}{2}$$

for $0<x<\pi$.

Subtracting from $f(x)$ the jump function $g(x)$, we get the function

$$\varphi(x)=x^2+\frac{1}{2}$$

which is continuous on the interval $[-\pi, \pi]$ (Fig. 48b). Since

$$\sigma_0(x)=\sum_{n=1}^{\infty}\frac{\sin nx}{n}$$

and

$$\sigma_0(x+\pi)=\sum_{n=1}^{\infty}\frac{\sin n(x+\pi)}{n}=\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nx$$

it follows that

$$g(x)=\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nx-\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin nx=$$

$$=\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n-1}{n}\sin nx=-\frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\sin(2k+1)x}{2k+1}$$

Hence

$$f(x)=g(x)+\frac{1}{2}+\frac{\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos nx$$

and the coefficients of the transformed Fourier series have the order of decay $O\left(\frac{1}{n^2}\right)$.

Note that if from $f(x)$ we isolate the jump function $g(x)$ to within the discontinuities of the derivative, then the remainder will be identically zero; that is, we get the exact sum of the series (10).

Note. The method of A. N. Krylov is also applicable to Fourier series of period $T=2l$. Indeed, let a function $f(x)$ be given in the main domain $a-l<x<a+l$. Performing the linear transformation

$$x=a+\frac{l}{\pi}t$$

we obtain the function $F(t)=f\left(a+\frac{l}{\pi}t\right)$ of period 2π defined in the standard domain $-\pi<t<\pi$.

6.5 TRIGONOMETRIC APPROXIMATION

Suppose we have a convergent trigonometric series

$$\sum_{n=0}^{\infty}(a_n\cos nx+b_n\sin nx)=S(x) \qquad (1)$$

whose sum is $S(x)$ and not known. It is required to compute this sum to a preassigned degree of accuracy.

It is obvious that the faster the coefficients a_n and b_n of (1) tend to zero, the smaller the number of terms we have to take to ensure the given accuracy. Therefore, it is best to accelerate the convergence of the series before computing the sum. The usual procedure is as follows: from the given series is extracted some trigonometric series, whose sum $g(x)$ is known, such that the remaining series

$$\sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx) \tag{2}$$

is more rapidly convergent than the original series.

If

$$g(x) = \sum_{n=0}^{\infty} (\overline{a}_n \cos nx + \overline{b}_n \sin nx)$$

then

$$S(x) = g(x) + \sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx) \tag{3}$$

where

$$\alpha_n = a_n - \overline{a}_n \qquad (n = 0,\ 1,\ 2,\ \ldots)$$

In the simplest cases, we can use the earlier considered expansions for constructing the function $g(x)$:

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n} = \sigma_0(x) = \frac{\pi - x}{2} \qquad (0 < x < 2\pi),$$

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} = -\sigma_1(x) = \frac{(\pi - x)^2}{4} - \frac{\pi^2}{12} \qquad (0 \leqslant x \leqslant 2\pi),$$

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^3} = -\sigma_2(x) = \frac{2\pi^2 x - 3\pi x^2 + x^3}{12} \qquad (0 \leqslant x \leqslant 2\pi),$$

. . . .

Also useful are the following expansions [7]

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n} = -\ln\left(2 \sin \frac{x}{2}\right) \qquad (0 < x < 2\pi),$$

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^2} = -\int_0^x \ln\left(2 \sin \frac{x}{2}\right) dx \quad (0 \leqslant x \leqslant 2\pi),$$

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^3} = \int_0^x dx \int_0^x \ln\left(2 \sin \frac{x}{2}\right) dx + \sum_{n=1}^{\infty} \frac{1}{n^3} \qquad (0 \leqslant x \leqslant 2\pi)$$

where $\sum_{n=1}^{\infty} \frac{1}{n^3} = 1.202056903\ldots.$

Example. Find the sum of the series

$$S(x)=\sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx$$

to within 0.001.

Solution. The coefficients $b_n=\frac{n}{n^2+1}$ of the series have an order of decay $O\left(\frac{1}{n}\right)$ since $\lim\limits_{n\to\infty}\left(b_n:\frac{1}{n}\right)=1$. Let us accelerate the convergence of the given series. It is clear that

$$\frac{n}{n^2+1}=\frac{n}{n^2}\left(\frac{1}{1+\frac{1}{n^2}}\right)=\frac{1}{n}\left(1-\frac{1}{n^2}+\frac{1}{n^4}-\dots\right)=\frac{1}{n}-\frac{1}{n^3}+\gamma_n$$

where

$$\gamma_n=\frac{n}{n^2+1}-\frac{1}{n}+\frac{1}{n^3}=\frac{1}{n^3(n^2+1)}$$

Then

$$\sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx=\sum_{n=1}^{\infty}\frac{\sin nx}{n}-\sum_{n=1}^{\infty}\frac{\sin nx}{n^3}+\sum_{n=1}^{\infty}\gamma_n\sin nx$$

But

$$\sum_{n=1}^{\infty}\frac{\sin nx}{n}=\sigma_0(x)\quad\text{and}\quad\sum_{n=1}^{\infty}\frac{\sin nx}{n^3}=-\sigma_2(x)$$

Thus

$$S(x)=\sigma_0(x)+\sigma_2(x)+\sum_{n=1}^{\infty}\gamma_n\sin nx$$

where $\gamma_n=\frac{1}{n^3(n^2+1)}=O\left(\frac{1}{n^5}\right)$.

Let N be the number of terms of the series $\sum\limits_{n=1}^{\infty}\gamma_n\sin nx$ which must be taken so that the remainder R_N satisfies the inequality

$$|R_N|=\left|\sum_{N+1}^{\infty}\gamma_n\sin nx\right|<0.001$$

Let us find the number N. We have

$$\left|\sum_{N+1}^{\infty}\frac{1}{n^3(n^2+1)}\sin nx\right|<\sum_{N+1}^{\infty}\frac{1}{n^5}<\int_N^{\infty}\frac{dx}{x^5}=\frac{1}{4N^4}$$

Solving the inequality $\frac{1}{4N^4}<0.001$ we find that $N=5$ suffices.

Hence, to the given accuracy we have

$$S(x)=\frac{\pi-x}{2}-\frac{2\pi^2x-3\pi x^2+x^3}{12}+\sum_{n=1}^{5}\frac{\sin nx}{n^3(n^2+1)} \qquad (0<x<\pi)$$

REFERENCES FOR CHAPTER 6

[1] *G. M. Fikhtengolts, Principles of Mathematical Analysis,* 1956, Vol.. II, Chapters XV and XXIV (in Russian).
[2] *A. Markov, Calculus of Finite Differences,* 1911, Chapter II (in Russian).
[3] *G. Salekhov, Calculation of Series,* 1955, Chapters I and III (in Russian).
[4] *Ya. S. Bezikovich, Calculus of Finite Differences,* 1939, Chapter IX (in Russian).
[5] *A. O. Gelfond, Calculus of Finite Differences,* 1952, Chapter IV (in Russian).
[6] *Vallée-Poussin, C. J de la, Cours d'Analyse Infinitesimale,* Vol. II, 1921.
[7] *G. P. Tolstov, Fourier Series,* 1951, Chapters I-V (in Russian).
[8] *A. N. Krylov, Lectures on Approximate Computations,* 1954, Chapter V (in Russian).
[9] *L. V. Kantorovich, V. I. Krylov, Approximate Methods of Higher Analysis,* 1949, Chapter I (in Russian).

Chapter 7

MATRIX ALGEBRA

7.1 BASIC DEFINITIONS

A set mn of numbers (real or complex) arranged in a rectangular array of m rows and n columns

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \qquad (1)$$

is called a *matrix* (of numbers). The rows and columns of (1) are termed the *lines* of the matrix.

The numbers a_{ij} $(i=1, 2, \dots, m;\ j=1, 2, \dots, n)$ that comprise the given matrix are called the *elements* (or *entries*) of the matrix. Here, the first subscript i denotes the number of the row of the element and the second subscript j denotes the number of its column.

A matrix, say (1), is often more compactly written as

$$A=[a_{ij}] \qquad (i=1, 2, \dots, m;\ j=1, 2, \dots, n)$$

or

$$A=[a_{ij}]_{m,\,n}$$

We say that the matrix A has dimensions $m\times n$, or that A is an m-by-n matrix (or an $m\times n$ matrix, or is of type $m\times n$).

If $m=n$, the matrix is called a *square matrix* of order n. If $m\neq n$, then it is a *rectangular matrix*. A $1\times n$ matrix is called a *row vector* and an $m\times 1$ matrix, a *column vector*. An ordinary number (scalar) may be regarded as a 1×1 matrix. A square matrix of the form

$$A=\begin{bmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ 0 & \alpha_2 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & \alpha_n \end{bmatrix} \qquad (2)$$

is termed *diagonal* and is briefly denoted as $[\alpha_1, \alpha_2, \dots, \alpha_n]$.

If $\alpha_i = 1$ $(i = 1, 2, \ldots, n)$, the matrix (2) is called a *unit matrix* and is denoted by the letter E (or I); thus,

$$E = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

Introducing a symbol called the *Kronecker delta,*

$$\delta_{ij} = \begin{cases} 0 & \text{if} \quad i \neq j, \\ 1 & \text{if} \quad i = j \end{cases}$$

we can write

$$E = [\delta_{ij}]$$

A matrix with all elements zero is called a *zero matrix* and is denoted by 0. To indicate the number of rows and columns of a zero matrix, one writes 0_{mn}.

With a square matrix $A = [a_{ij}]_{n,\,n}$ is associated a so-called *determinant*:

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

These are two distinct concepts: a matrix is an **ordered set** of numbers written in the form of a rectangular array; its determinant, $\det A$, is a number that is determined by means of specific rules, namely:

$$\det A = \sum_{(\alpha_1,\, \alpha_2,\, \ldots,\, \alpha_n)} (-1)^{\varkappa} a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n} \tag{3}$$

where the summation (3) is taken over all permutations $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of the elements $1, 2, \ldots, n$, and, consequently, contains $n!$ summands; $\varkappa = 0$ if the permutation is even and $\varkappa = 1$ if the permutation is odd.

7.2 OPERATIONS INVOLVING MATRICES

A. EQUALITY OF MATRICES

Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are considered equal, $A = B$, if they have the same dimensions, which is to say, if they have the same number of rows and columns, and the corresponding elements are equal; thus,

$$a_{ij}=b_{ij}$$

B. THE SUM AND DIFFERENCE OF MATRICES

The *sum of two matrices* $A=[a_{ij}]$ and $B=[b_{ij}]$ of the same dimensions is a matrix $C=[c_{ij}]$ of the same dimensions with elements c_{ij} equal to the sums of the corresponding elements a_{ij} and b_{ij} of the matrices A and B; that is, $c_{ij}=a_{ij}+b_{ij}$. Thus,

$$A+B=\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2n}+b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{bmatrix}$$

The following properties are derived directly from the definition of a matrix sum:

(1) $A+(B+C)=(A+B)+C$,
(2) $A+B=B+A$,
(3) $A+0=A$.

The *difference of two matrices* is defined analogously:

$$A-B=\begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \dots & a_{1n}-b_{1n} \\ a_{21}-b_{21} & a_{22}-b_{22} & \dots & a_{2n}-b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}-b_{m1} & a_{m2}-b_{m2} & \dots & a_{mn}-b_{mn} \end{bmatrix}$$

C. MULTIPLICATION OF A MATRIX BY A SCALAR

The *product of a matrix* $A=[a_{ij}]$ by a scalar α (or the product of a scalar by a matrix A) is a *matrix* whose elements are obtained by multiplying all the elements of A by the scalar α; that is,

$$A\alpha=\alpha A=\begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{bmatrix}$$

From the definition of the product of a scalar by a matrix follow directly the properties:

(1) $1A=A$,
(2) $0A=0$,
(3) $\alpha(\beta A)=(\alpha\beta)A$,

(4) $(\alpha+\beta)A=\alpha A+\beta A$,
(5) $\alpha(A+B)=\alpha A+\alpha B$.

Here, A and B are matrices, and α and β are scalars.

Note that if matrix A is square of order n, then

$$\det \alpha A=\alpha^n \det A$$

The matrix

$$-A=(-1)A$$

is called the *negative (additive inverse)* of A. It is easy to see that if A and B have the same dimensions, then

$$A-B=A+(-B)$$

D. MULTIPLICATION OF MATRICES

Suppose

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

and

$$B=\begin{bmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ b_{p1} & b_{p2} & \dots & b_{pq} \end{bmatrix}$$

are matrices of dimensions $m\times n$ and $p\times q$, respectively. If the number of columns of A is equal to the number of rows of B, that is,

$$n=p \tag{1}$$

then for these matrices is defined a *product*, matrix C of dimensions $m\times q$:

$$C=\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ c_{m1} & c_{m2} & \dots & c_{mq} \end{bmatrix}$$

where

$$c_{ij}=a_{i1}b_{1j}+a_{i2}b_{2j}+\dots+a_{in}b_{nj} \qquad (i=1,2,\dots,m;\ j=1,2,\dots,q)$$

From the definition follows the rule for multiplication of matrices: *to obtain the element in the* i*th row and* j*th column of the pro-*

duct of two matrices, multiply the elements of the i*th row of the first matrix by the corresponding elements of the* j*th column of the second and add the products.*

The product AB is meaningful if and only if the matrix A contains as many elements in the rows as there are elements in the columns of matrix B. In particular, it is only possible to multiply square matrices of the same order.

Example 1.

$$A=\begin{bmatrix}3 & 2 & 8 & 1\\ 1 & -4 & 0 & 3\end{bmatrix},$$

$$B=\begin{bmatrix}2 & -1\\ 1 & -3\\ 0 & 1\\ 3 & 1\end{bmatrix},$$

$$AB=$$

$$=\begin{bmatrix}3\cdot 2+2\cdot 1+8\cdot 0+1\cdot 3 & 3\cdot(-1)+2\cdot(-3)+8\cdot 1+1\cdot 1\\ 1\cdot 2+(-4)\cdot 1+0\cdot 0+3\cdot 3 & 1\cdot(-1)+(-4)\cdot(-3)+0\cdot 1+3\cdot 1\end{bmatrix}=$$

$$=\begin{bmatrix}11 & 0\\ 7 & 14\end{bmatrix}$$

Example 2.

$$\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix}\cdot\begin{bmatrix}1\\ 2\\ 3\end{bmatrix}=\begin{bmatrix}1\cdot 1+2\cdot 2+3\cdot 3\\ 4\cdot 1+5\cdot 2+6\cdot 3\\ 7\cdot 1+8\cdot 2+9\cdot 3\end{bmatrix}=\begin{bmatrix}14\\ 32\\ 50\end{bmatrix}$$

A matrix product has the following properties:

$$(1)\ A(BC)=(AB)C,\quad (3)\ (A+B)C=AC+BC,$$
$$(2)\ \alpha(AB)=(\alpha A)B,\quad (4)\ C(A+B)=CA+CB$$

where A, B and C are matrices and α is a scalar.

Equations (1) to (4) are to be understood in the sense that if one of their members exists, then the other member also exists, and they are equal.

The product of two matrices is generally noncommutative, $AB\neq BA$, as witness the examples.

Example 3.

$$A=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix},\qquad B=\begin{bmatrix}5 & 6\\ 7 & 8\end{bmatrix}$$

Then

$$AB=\begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix},\qquad BA=\begin{bmatrix}23 & 34\\ 31 & 46\end{bmatrix}$$

which is to say, $AB\neq BA$.

What is more, it may happen that the product of two matrices in a given order is meaningful while the product of the same matrices in the reverse order is quite meaningless.

For example, if

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}, \quad B=\begin{bmatrix}3 & 2 & 1\\ 2 & 1 & 3\\ 4 & 3 & 0\end{bmatrix}$$

then

$$AB=\begin{bmatrix}19 & 13 & 7\\ 46 & 31 & 19\end{bmatrix}$$

but BA does not exist.

When $AB=BA$, the matrices A and B are said to be *commutative*. Thus, for example, as is readily seen, the unit matrix E is commutative with any square matrix A of the same order, and

$$AE=EA=A$$

Thus the unit matrix E *plays the role of identity (unity)* in multiplication.

If A and B are square matrices of the same order, then

$$\det(AB)=\det(BA)=\det A\cdot\det B$$

This formula follows from the rule for multiplying determinants.

To illustrate, for the matrices given in Example 3 we have

$$\begin{vmatrix}19 & 22\\ 43 & 50\end{vmatrix}=\begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}$$

and

$$\begin{vmatrix}23 & 34\\ 31 & 46\end{vmatrix}=\begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}$$

7.3 THE TRANSPOSE OF A MATRIX

If in an $m\times n$ matrix

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{m1} & a_{m2} & \dots & a_{mn}\end{bmatrix}$$

we replace the rows by the columns, we get what is called the *transpose* of A:

$$A'=A^T=\begin{bmatrix}a_{11} & a_{21} & \dots & a_{m1}\\ a_{12} & a_{22} & \dots & a_{m2}\\ \cdot & \cdot & \cdot & \cdot\\ a_{1n} & a_{2n} & \dots & a_{mn}\end{bmatrix}$$

of dimensions $n \times m$. In particular, the transpose of the row vector

$$\boldsymbol{a} = [a_1 \ a_2 \ \dots \ a_n]$$

is the column vector

$$\boldsymbol{a}' = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix}$$

The transpose of a matrix has the following properties:

(1) the transpose of a transpose is the original matrix:

$$A'' = (A')' = A$$

(2) the transpose of a sum is equal to the sum of the transposed matrices of the summands, i. e.,

$$(A+B)' = A' + B'$$

(3) the transpose of a product is equal to the product of the transposes of the factors taken in reverse order:

$$(AB)' = B'A'$$

Indeed, the element of the ith row and jth column of the matrix $(AB)'$ is equal to the element of the jth row and ith column of the matrix AB:

$$a_{j1}b_{1i} + a_{j2}b_{2i} + \dots + a_{jn}b_{ni}$$

This expression is obviously the sum of the products of the elements of the ith row of matrix B' by the corresponding elements of the jth column of matrix A'; that is to say, it is equal to the common element of the matrix $B'A'$. If A is square, then, clearly,.

$$\det A' = \det A$$

The matrix $A = [a_{ij}]$ is called a *symmetric matrix* if it coincides with its transpose, that is, if

$$A' = A \tag{1}$$

From equation (1) it follows that: (1) a symmetric matrix is square ($m = n$) and (2) the elements symmetric about the principal diagonal are equal, or

$$a_{ji} = a_{ij}$$

The product

$$C = AA'$$

is obviously a symmetric matrix, since

$$C' = (AA')' = (A')' A' = AA' = C$$

For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1^2 + 2^2 + 3^2 & 1\cdot 4 + 2\cdot 5 + 3\cdot 6 \\ 4\cdot 1 + 5\cdot 2 + 6\cdot 3 & 4^2 + 5^2 + 6^2 \end{bmatrix} = $$
$$= \begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix}$$

7.4 THE INVERSE MATRIX

Definition 1. *The inverse of a given matrix* is a matrix such that, when multiplied on the right (postmultiplied) or on the left (premultiplied) by the given matrix, yields the unit matrix.

We denote the inverse of matrix A by A^{-1}. Then, by definition, we have

$$AA^{-1} = A^{-1}A = E \qquad (1)$$

where E is the unit matrix.

Finding the inverse of a given matrix is called *matrix inversion.*

Definition 2. A square matrix is termed *nonsingular* if the determinant is different from zero, otherwise it is called a *singular matrix.*

Theorem. *Every nonsingular matrix has an inverse.*

Proof. Suppose we have a nonsingular matrix of order n:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

where $\det A = \Delta \neq 0$.

We form the so-called *adjoint* of matrix A:

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \cdot & \cdot & \cdot & \cdot \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix} \qquad (2)$$

where A_{ij} are the cofactors (signed minors) of the corresponding elements a_{ij} $(i, j = 1, 2, \dots, n)$.

Note that the cofactors of the elements of the rows lie in the corresponding columns, that is, the transpose operation is performed.

Divide all the elements of the last matrix by the value of the determinant of A (by Δ that is):

$$A^* = \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \dots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \dots & \frac{A_{n2}}{\Delta} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \dots & \frac{A_{nn}}{\Delta} \end{bmatrix} \tag{3}$$

We will prove that the matrix A^* is the required inverse: $A^* = A^{-1}$.

As we know, (1) the sum of the products of the elements of a certain line (row or column) of the determinant by the cofactors of the elements is equal to the determinant, and (2) the sum of the products of the elements of a line of the determinant by the cofactors of the corresponding elements of a parallel line (row or column) is equal to zero; thus

$$\sum_{k=1}^{n} a_{ik} A_{jk} = \delta_{ij} \Delta \tag{4}$$

and

$$\sum_{k=1}^{n} a_{ki} A_{kj} = \delta_{ij} \Delta \tag{4'}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for} \quad i = j, \\ 0 & \text{for} \quad i \neq j \end{cases}$$

Using these properties, form the product AA^* to get

$$AA^* = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \dots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \dots & \frac{A_{n}}{\Delta} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \dots & \frac{A_{nn}}{\Delta} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix} = E \tag{5}$$

Thus, $AA^* = E$.

Formula (5) can be derived faster if we use the compact notation

$$A=[a_{ij}] \quad \text{and} \quad A^*=\left[\frac{A_{ji}}{\Delta}\right]$$

Taking relation (4) into account, we obtain

$$AA^*=\left[\sum_{k=1}^{n} a_{ik}\frac{A_{jk}}{\Delta}\right]=[\delta_{ij}]=E.$$

Similarly, we see that $A^*A=E$.

Hence, $A^*=A^{-1}$, or

$$A^{-1}=\frac{1}{\Delta}[A_{ji}] \tag{6}$$

where

$$\Delta=\det A$$

Note 1. The inverse A^{-1} of a given matrix A is unique. What it more, every right inverse (left inverse) of A coincides with its inverse A^{-1} (if such exists).

Indeed, if

$$AB=E$$

then, premultiplying this equation by A^{-1}, we get

$$A^{-1}AB=A^{-1}E$$

or

$$B=A^{-1}$$

We similarly prove that if

$$CA=E$$

then $C=A^{-1}$.

Therefore, when verifying relation (1) one equation is sufficient.

Note 2. A singular square matrix does not have an inverse. True enough: since the matrix A is singular,

$$\det A=0$$

From (1) we have

$$\det A\cdot\det A^{-1}=\det E=1$$

or

$$0=1\ (?!)$$

which is impossible. The assertion is proved.

Example. Find the inverse of the matrix

$$A=\begin{bmatrix} 1 & 2 & 3 \\ -2 & -4 & -5 \\ 3 & 5 & 6 \end{bmatrix}$$

Solution. Since the determinant

$$\Delta=\begin{vmatrix}1 & 2 & 3\\ -2 & -4 & -5\\ 3 & 5 & 6\end{vmatrix}=\begin{vmatrix}1 & 2 & 3\\ 0 & 0 & 1\\ 0 & -1 & -3\end{vmatrix}=1\neq 0$$

the matrix A is nonsingular.

Form the adjoint matrix

$$\tilde{A}=\begin{bmatrix}1 & 3 & 2\\ -3 & -3 & -1\\ 2 & 1 & 0\end{bmatrix}$$

Divide all elements of $\tilde{A}$ by $\Delta=1$ to get

$$A^{-1}=\begin{bmatrix}1 & 3 & 2\\ -3 & -3 & -1\\ 2 & 1 & 0\end{bmatrix}$$

The reader is advised to verify that we indeed have

$$AA^{-1}=E$$

Below are some of the basic properties of an inverse matrix.

1. *The determinant of an inverse matrix is equal to the reciprocal of the determinant of the original matrix.* Suppose

$$A^{-1}A=E$$

Taking into consideration that the determinant of a product of two square matrices is equal to the product of the determinants of the matrices, we get

$$\det A^{-1}\det A=\det E=1$$

Hence

$$\det A^{-1}=\frac{1}{\det A}$$

2. *The inverse of a product of square matrices is equal to the product of the inverses of the factors taken in reverse order:*

$$(AB)^{-1}=B^{-1}A^{-1}$$

Indeed,

$$AB(B^{-1}A^{-1})=A(BB^{-1})A^{-1}=AEA^{-1}=AA^{-1}=E$$

and

$$(B^{-1}A^{-1})AB=B^{-1}(A^{-1}A)B=B^{-1}EB=B^{-1}B=E$$

Hence $B^{-1}A^{-1}$ is the inverse of AB.

In the more general case,

$$(A_1A_2\dots A_p)^{-1}=A_p^{-1}A_{p-1}^{-1}\dots A_1^{-1}$$

3. *The transpose of an inverse is equal to the inverse of the transpose of the given matrix:*

$$(A^{-1})'=(A')^{-1}$$

Taking transposes of $A^{-1}A=E$, we get

$$(A^{-1}A)'=A'(A^{-1})'=E'=E$$

Whence, premultiplying the last equation by the matrix $(A')^{-1}$, we obtain

$$(A')^{-1}A'(A^{-1})'=(A')^{-1}E$$

or

$$(A^{-1})'=(A')^{-1}$$

which is what we set out to prove.

Note. The matrix equations

$$AX=B \quad \text{and} \quad YA=B$$

are easily solved by means of an inverse matrix.

If $\det A\neq 0$, then

$$X=A^{-1}B \quad \text{and} \quad Y=BA^{-1}$$

7.5 POWERS OF A MATRIX

Let A be a square matrix. If p is a natural number, then put

$$\underbrace{AA\dots A}_{p\text{ times}}=A^p$$

We also agree that $A^0=E$, where E is the unit matrix. If matrix A is nonsingular, we can introduce a negative power and define it by the relation

$$A^{-p}=(A^{-1})^p$$

The ordinary rules hold for powers of matrices with integral exponents:

(1) $A^pA^q=A^{p+q}$,

(2) $(A^p)^q=A^{pq}$.

It is obviously impossible to raise a nonsquare matrix to a power.

Example 1. Let

$$A=\begin{bmatrix}\alpha_1 & 0 & \dots & 0\\ 0 & \alpha_2 & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & \alpha_n\end{bmatrix}$$

Then

$$A^p=\begin{bmatrix}\alpha_1^p & 0 & \dots & 0\\ 0 & \alpha_2^p & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & \alpha_n^p\end{bmatrix}$$

Example 2. Find.

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}^2$$

Solution. We have

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}^2=\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

If A and B are square matrices of one and the same order, and $AB=BA$, then the binomial formula holds true:

$$(A+B)^p=\sum_{k=0}^{p}C_p^k A^k B^{p-k}$$

7.6 RATIONAL FUNCTIONS OF A MATRIX

Suppose

$$X=\begin{bmatrix}x_{11} & x_{12} & \dots & x_{1n}\\ x_{21} & x_{22} & \dots & x_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ x_{n1} & x_{n2} & \dots & x_{nn}\end{bmatrix}$$

is an arbitrary square matrix of order n. By analogy with the formulas of elementary algebra we determine the integral rational functions of the matrix X:

$$P(X)=A_0X^m+A_1X^{m-1}+\dots+A_mE \text{ (right polynomial)}$$
$$\tilde{P}(X)=X^mA_0+X^{m-1}A_1+\dots+EA_m \text{ (left polynomial)}$$

where $A_\nu (\nu = 0, 1, \ldots, m)$ are $m \times n$ or, respectively, $n \times m$ matrices and E is a unit matrix of order n.

Generally speaking, $P(X) \neq \tilde{P}(X)$.

It is also possible to introduce *fractional rational functions* of the matrix X, defining them by the formulas

$$R_1(X) = P(X)[Q(X)]^{-1}$$

and

$$R_2(X) = [Q(X)]^{-1} P(X)$$

where $P(X)$ and $Q(X)$ are matrix polynomials and $\det [Q(X)] \neq 0$.

Example. Let

$$P(X) = X^2 + \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} X - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where X is a variable matrix of order two. Find $P\left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right)$.

Solution. We have

$$P\left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}^2 + \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}$$

7.7 THE ABSOLUTE VALUE AND NORM OF A MATRIX

The inequality

$$A \leqslant B \tag{1}$$

between two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ of the same size means that

$$a_{ij} \leqslant b_{ij} \tag{2}$$

In this sense, just any two matrices are not always comparable.

We use the term *absolute value* (*modulus*) of a matrix $A = [a_{ij}]$ to mean the matrix

$$|A| = [|a_{ij}|]$$

where $|a_{ij}|$ are the moduli of the elements of A.

If A and B are matrices for which the operations $A + B$ and AB are meaningful, then

(a) $|A + B| \leqslant |A| + |B|$,
(b) $|AB| \leqslant |A| \cdot |B|$,
(c) $|\alpha A| = |\alpha| \, |A|$

where α is a scalar.

In particular, we get

$$|A^p| \leqslant |A|^p$$

where p is a natural number.

By the *norm* of a matrix $A=[a_{ij}]$ we mean a real number $\|A\|$ which satisfies the following conditions:

(a) $\|A\| \geqslant 0$ with $\|A\|=0$ if and only if $A=0$,

(b) $\|\alpha A\|=|\alpha|\,\|A\|$ (α a scalar) and, in particular, $\|-A\|=\|A\|$,

(c) $\|A+B\| \leqslant \|A\|+\|B\|$,

(d) $\|AB\| \leqslant \|A\| \cdot \|B\|$

(A and B are matrices for which the corresponding operations are meaningful.) As a particular instance, for the square matrix we have

$$\|A^p\| \leqslant \|A\|^p$$

where p is a natural number.

We note one more important inequality between the norms of matrices A and B of the same size. Using Condition (c), we have

$$\|B\|=\|A+(B-A)\| \leqslant \|A\|+\|B-A\|$$

whence

$$\|A-B\|=\|B-A\| \geqslant \|B\|-\|A\|$$

Similarly

$$\|A-B\| \geqslant \|A\|-\|B\|$$

hence

$$\|A-B\| \geqslant |\,\|B\|-\|A\|\,|$$

We call the norm *canonical* if the following two conditions hold true as well:

(e) if $A=[a_{ij}]$, then

$$|a_{ij}| \leqslant \|A\|$$

and for the scalar matrix $A=[a_{11}]$ we have $\|A\|=|a_{11}|$,

(f) from the inequality $|A| \leqslant |B|$ (A and B are matrices) follows the inequality

$$\|A\| \leqslant \|B\|$$

In particular, $\|A\|=\|\,|A|\,\|$.

In the sequel, for any matrix $A=[a_{i}\ ]$ of arbitrary dimensions we will consider mainly three norms that are easily computed:

(1) $\|A\|_m=\max\limits_{i} \sum\limits_{j} |a_{ij}|$ (the *m-norm*),

(2) $\|A\|_l=\max\limits_{j} \sum\limits_{i} |a_{ij}|$ (the *l-norm*),

(3) $\|A\|_k=\sqrt{\sum\limits_{i,j} |a_{ij}|^2}$ (the *k-norm*).

Example. Suppose

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix}$$

We have

$$\|A\|_m=\max(1+2+3,\ 4+5+6,\ 7+8+9)=\max(6,\ 15,\ 24)=24,$$
$$\|A\|_l=\max(1+4+7,\ 2+5+8,\ 3+6+9)=\max(12,\ 15,\ 18)=18,$$
$$\|A\|_k=\sqrt{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2}=$$
$$=\sqrt{1+4+9+16+25+36+49+64+81}=\sqrt{285}\approx 16.9$$

In particular, for the vector

$$\boldsymbol{x}=\begin{bmatrix}x_1\\ x_2\\ \cdot\\ \cdot\\ \cdot\\ x_n\end{bmatrix}$$

these norms have the following values:

$$\|\boldsymbol{x}\|_m=\max_i |x_i|,$$
$$\|\boldsymbol{x}\|_l=|x_1|+|x_2|+\ldots+|x_n|,$$
$$\|\boldsymbol{x}\|_k=|\boldsymbol{x}|=\sqrt{|x_1|^2+|x_2|^2+\ldots+|x_n|^2}$$

(the *absolute value of the vector*). If the components of the vector are real, then we simply have

$$\|\boldsymbol{x}\|_k=\sqrt{x_1^2+x_2^2+\ldots+x_n^2}$$

Let us verify the Conditions (a) to (d) for the norms $\|A\|_m$, $\|A\|_l$ and $\|A\|_k$.

It is immediately obvious that the Conditions (a) and (b) are fulfilled. Let us assure ourselves that Condition (c) is fulfilled as well. Let $A=[a_{ij}]$ and $B=[b_{ij}]$, A and B being of the same size. We have

$$\|A+B\|_m=\max_i\sum_j|a_{ij}+b_{ij}|\leqslant\max_i\left\{\sum_j|a_{ij}|+\sum_j|b_{ij}|\right\}\leqslant$$
$$\leqslant\max_i\sum_j|a_{ij}|+\max_i\sum_j|b_{ij}|=\|A\|_m+\|B\|_m$$

Similarly

$$\|A+B\|_l\leqslant\|A\|_l+\|B\|_l$$

Furthermore,

$$\| A+B \|_k = \sqrt{\sum_{i,j} | a_{ij}+b_{ij} |^2} \leqslant \sqrt{\sum_{i,j} | a_{ij} |^2 + \sum_{i,j} | b_{ij} |^2 + 2 \sum_{i,j} | a_{ij} | | b_{ij} |}$$

Applying the familiar Cauchy inequality[1]

$$\sum_{i,j} | a_{ij} | | b_{ij} | \leqslant \sqrt{\sum_{i,j} | a_{ij} |^2} \cdot \sqrt{\sum_{i,j} | b_{ij} |^2}$$

---

[1] The proof of the *Cauchy inequality* follows:

$$\left| \sum_{s=1}^{n} a_s b_s \right|^2 \leqslant \sum_{s=1}^{n} | a_s |^2 \cdot \sum_{s=1}^{n} | b_s |^2$$

where a_s and b_s $(s=1, 2, \ldots, n)$ are arbitrary complex numbers. Let λ be a real variable. We consider the obvious inequality

$$\sum_{s=1}^{n} | a_s \lambda + b_s e^{i\varphi_s} |^2 \geqslant 0 \tag{$*$}$$

where φ_s are some real numbers. Denoting by $\bar{a}_s$ and $\bar{b}_s$ the conjugates of a_s and b_s, we have

$$\begin{aligned} | a_s \lambda + b_s e^{i\varphi_s} |^2 &= (a_s \lambda + b_s e^{i\varphi_s}) (\bar{a}_s \lambda + \bar{b}_s e^{-i\varphi_s}) = \\ &= a_s \bar{a}_s \lambda^2 + (a_s \bar{b}_s e^{-i\varphi_s} + \bar{a}_s b_s e^{i\varphi_s}) \lambda + b_s \bar{b}_s = \\ &= | a_s |^2 \lambda^2 + 2 \operatorname{Re} (a_s \bar{b}_s e^{-i\varphi_s}) \lambda + | b_s |^2 \end{aligned}$$

The inequality $(*)$ then becomes

$$\lambda^2 \sum_{s=1}^{n} | a_s |^2 + 2\lambda \sum_{s=1}^{n} \operatorname{Re} (a_s \bar{b}_s e^{-i\varphi_s}) + \sum_{s=1}^{n} | b_s |^2 \geqslant 0$$

If we put

$$\varphi_s = \arg (a_s \bar{b}_s)$$

then

$$\begin{aligned} \operatorname{Re} (a_s \bar{b}_s e^{-i\varphi_s}) &= \operatorname{Re} \{ | a_s \bar{b}_s | e^{i \arg (a_s \bar{b}_s)} \cdot e^{-i \arg (a_s \bar{b}_s)} \} = \\ &= \operatorname{Re} \{ | a_s \bar{b}_s | \} = | a_s \bar{b}_s | = | a_s b_s | \end{aligned}$$

and, consequently,

$$\lambda^2 \sum_{s=1}^{n} | a_s |^2 + 2\lambda \sum_{s=1}^{n} | a_s b_s | + \sum_{s=1}^{n} | b_s |^2 \geqslant 0.$$

Since, by virtue of the inequality $(*)$, the left member of this inequality is nonnegative for arbitrary real λ, the corresponding quadratic equation cannot have distinct real roots. Therefore the discriminant of the equation is

$$\left\{ \sum_{s=1}^{n} | a_s b_s | \right\}^2 - \sum_{s=1}^{n} | a_s |^2 \cdot \sum_{s=1}^{n} | b_s |^2 \leqslant 0$$

that is

$$\left\{ \sum_{s=1}^{n} | a_s b_s | \right\}^2 \leqslant \sum_{s=1}^{n} | a_s |^2 \cdot \sum_{s=1}^{n} | b_s |^2$$

(see over)

we get

$$\|A+B\|_k \leqslant \sqrt{\sum_{i,j}|a_{ij}|^2}+\sqrt{\sum_{i,j}|b_{ij}|^2}=\|A\|_k+\|B\|_k$$

The Condition (c) is thus fulfilled for all three norms.

Let us now verify it for Condition (d). Suppose matrix $A=[a_{ij}]$ is of dimensions $m'\times n'$ and matrix $B=[b_{ij}]$ is of dimensions $m''\times n''$. For the possibility of multiplication of the first matrix by the second, it is necessary that $m''=n'$, and the matrix AB will have the dimensions $m'\times n''$.

We have

$$\|AB\|_m=\max_i\sum_{j=1}^{n''}\left|\sum_{s=1}^{n'}a_{is}a_{sj}\right|\leqslant$$
$$\leqslant\max_i\left\{\sum_{j=1}^{n''}\sum_{s=1}^{n'}|a_{is}||b_{sj}|\right\}=$$
$$=\max_i\left\{\sum_{s=1}^{n'}|a_{is}|\sum_{j=1}^{n''}|b_{sj}|\right\}\leqslant$$
$$\leqslant\max_i\left\{\sum_{s=1}^{n'}|a_{is}|\cdot\|B\|_m\right\}=$$
$$=\max_i\left\{\sum_{s=1}^{n'}|a_{is}|\right\}\cdot\|B\|_m=\|A\|_m\cdot\|B\|_m$$

Similarly

$$\|AB\|_l=\max_j\sum_{i=1}^{m'}\left|\sum_{s=1}^{n'}a_{is}b_{sj}\right|\leqslant$$
$$\leqslant\max_j\left\{\sum_{i=1}^{m'}\sum_{s=1}^{n'}|a_{is}||b_{sj}|\right\}=$$
$$=\max_j\left\{\sum_{s=1}^{n'}|b_{sj}|\sum_{i=1}^{m'}|a_{is}|\right\}\leqslant$$
$$\leqslant\max_j\left\{\sum_{s=1}^{n'}|b_{sj}|\cdot\|A\|_l\right\}=$$
$$=\|A\|_l\cdot\max_j\sum_{s=1}^{n'}|b_{sj}|=\|A\|_l\cdot\|B\|_l$$

---

whence, all the more so,

$$\left|\sum_{s=1}^{n}a_sb_s\right|^2\leqslant\left\{\sum_{s=1}^{n}|a_sb_s|\right\}^2\leqslant\sum_{s=1}^{n}|a_s|^2\cdot\sum_{s=1}^{n}|b_s|^2$$

If the numbers a_s and b_s are real, then we simply get

$$\left(\sum_{s=1}^{n}a_sb_s\right)^2\leqslant\sum_{s=1}^{n}a_s^2\cdot\sum_{s=1}^{n}b_s^2$$

Furthermore,

$$\| AB \|_k = \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left| \sum_{s=1}^{n'} a_{is} b_{sj} \right|^2} \leqslant \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left\{ \sum_{s=1}^{n'} | a_{is} | | b_{sj} | \right\}^2}$$

Applying the Cauchy inequality and taking into account that $m'' = n'$, we obtain

$$\| AB \|_k \leqslant \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left\{ \sum_{s=1}^{n'} | a_{is} |^2 \cdot \sum_{t=1}^{m''} | b_{tj} |^2 \right\}} =$$

$$= \sqrt{\sum_{i=1}^{m'} \sum_{s=1}^{n'} | a_{is} |^2 \cdot \sum_{t=1}^{m''} \sum_{j=1}^{n''} | b_{tj} |^2} = \sqrt{\| A \|_k^2 \cdot \| B \|_k^2} = \| A \|_k \cdot \| B \|_k$$

Hence, Condition (d) is fulfilled for the norms under consideration.

We will now show that the norms $\| A \|_m$, $\| A \|_l$ and $\| A \|_k$ are canonical.

If a_{pq} is the largest, in modulus, element of the matrix $A = [a_{ij}]$ of dimensions $m' \times n'$, then we obviously have

$$\| a \|_m \geqslant | a_{p1} | + \ldots + | a_{pq} | + \ldots + | a_{pn'} | \geqslant | a_{pq} |,$$

$$\| A \|_l \geqslant | a_{1q} | + \ldots + | a_{pq} | + \ldots + | a_{m'q} | \geqslant | a_{pq} |$$

and

$$\| A \|_k = \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n'} | a_{ij} |^2} \geqslant | a_{pq} |$$

Thus

$$| a_{ij} | \leqslant | a_{pq} | \leqslant \| A \|_s \qquad (s = m,\ l,\ k)$$

Besides, if $A = [a_{11}]$, then

$$\| A \|_m = \| A \|_l = \| A \|_k = | a_{11} |$$

Furthermore, if $| A | \leqslant | B |$, where $A = [a_{ij}]$ and $B = [b_{ij}]$, then $| a_{ij} | \leqslant | b_{ij} |$. From the definition of the norms $\| A \|_m$, $\| A \|_l$ and $\| A \|_k$ it is obvious that the inequalities

$$\| A \|_s \leqslant \| B \|_s \ (s = m,\ l,\ k)$$

hold true.

Besides, for any one of the norms we have

$$\| A \|_s = \| \, | A | \, \|_s \qquad (s = m,\ l,\ k)$$

Thus, Condition (f) is also fulfilled.

We have thus proved that the norms $\| A \|_m$, $\| A \|_l$ and $\| A \|_k$ are canonical.

Note that if matrix E is a unit matrix of order n, then

$$\|E\|_m = \|E\|_l = 1$$

and

$$\|E\|_k = \sqrt{n}$$

7.8 THE RANK OF A MATRIX

Suppose we have a rectangular matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

If in this matrix we choose in arbitrary fashion k rows and k columns, where $k \leqslant \min(m, n)$, then the elements at the intersections of these rows and columns form a square matrix of order k. The determinant of this latter matrix is called a *minor of kth order* of the matrix A.

Definition. *The rank of a matrix is the order of the nonvanishing minor of greatest order of the matrix.* In other words, a matrix A has rank r if:

(1) there is at least one minor of rth order different from zero;

(2) all minors of A of order $r+1$ and higher are equal to zero.

The rank of a zero matrix, that is, one consisting of zeros, is taken to be zero. The difference between the smallest of the numbers m and n and the rank of the matrix is termed the *nullity* of the matrix. If the nullity is zero, then the rank of the matrix is the largest of possible ranks for the given dimensions.

When determining the rank of a matrix it is useful to abide by the following rules:

(1) go from minors of small orders (beginning with minors of order one, that is, with the elements of the matrix) to minors of larger orders;

(2) suppose we have found a nonzero minor D of order r, then we have only to compute the minors of order $(r+1)$ bordering the minor D. If all these minors are zero, then the rank of the matrix is r; but if even one of them is nonzero, then this operation has to be applied to it, and then the rank of the matrix is definitely greater than r.

Example. Find the rank of the matrix

$$\begin{bmatrix} 2 & -4 & 3 & 1 & 0 \\ 1 & -2 & 1 & -4 & 2 \\ 0 & 1 & -1 & 3 & 1 \\ 4 & -7 & 4 & -4 & 5 \end{bmatrix}$$

Solution. The second-order minor in the upper left-hand corner of the matrix is zero. However, the matrix contains nonzero minors of the second order, as for example

$$D=\begin{vmatrix} -4 & 3 \\ -2 & 1 \end{vmatrix} \neq 0$$

and the third-order minor bordering it:

$$D'=\begin{vmatrix} 2 & -4 & 3 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{vmatrix} = 1$$

and both fourth-order minors bordering D' are equal to zero:

$$\begin{vmatrix} 2 & -4 & 3 & 1 \\ 1 & -2 & 1 & -4 \\ 0 & 1 & -1 & 3 \\ 4 & -7 & 4 & -4 \end{vmatrix} = 0, \qquad \begin{vmatrix} 2 & -4 & 3 & 0 \\ 1 & -2 & 1 & 2 \\ 0 & 1 & -1 & 1 \\ 4 & -7 & 4 & 5 \end{vmatrix} = 0$$

Hence the rank of the matrix is three, and the nullity is $4-3=1$.

7.9 THE LIMIT OF A MATRIX

Suppose we have a sequence of matrices

$$A_k=[a_{ij}^{(k)}] \qquad (k=1,\ 2,\ \ldots) \tag{1}$$

of the same dimensions $m\times n\,(i=1,\ 2,\ \ldots,\ m;\ j=1,\ 2,\ \ldots,\ n)$. By the *limit* of the sequence of matrices A_k is meant the matrix

$$A=\lim_{k\to\infty} A_k=[\lim_{k\to\infty} a_{ij}^{(k)}] \tag{2}$$

A sequence of matrices having a limit is called a *convergent sequence.*

Lemma 1. *For a sequence of matrices* $A_k\,(k=1,\ 2,\ \ldots)$ *to converge to the matrix* A *it is necessary and sufficient that*

$$\|A-A_k\|\longrightarrow 0 \text{ as } k\longrightarrow\infty \tag{3}$$

where $\|A\|$ *is any canonical norm of* A. *Here*

$$\lim_{k\to\infty}\|A_k\|=\|A\|$$

Indeed, if

$$A_k\longrightarrow A=[a_{ij}]$$

then

$$|a_{ij}-a_{ij}^{(k)}|<\varepsilon \quad \text{for} \quad k>N(\varepsilon)$$

whence

$$|A-A_k|<\varepsilon I$$

where I is an $m\times n$ matrix all elements of which are unity. By the properties of the norm we have

$$\|A-A_k\|\leqslant\varepsilon\|I\| \quad \text{for} \quad k>N(\varepsilon)$$

hence,

$$\lim_{k\to\infty}\|A-A_k\|=0 \tag{4}$$

Conversely, let Condition (3) be valid. Then for $k>N(\varepsilon)$ we have

$$|a_{ij}-a_{ij}^{(k)}|\leqslant\|A-A_k\|<\varepsilon$$

and hence

$$\lim_{k\to\infty}a_{ij}^{(k)}=a_{ij}$$

that is,

$$\lim_{k\to\infty}A_k=A$$

Besides, if $A_k\longrightarrow A$, then we have

$$|\|A\|-\|A_k\||\leqslant\|A-A_k\|\longrightarrow 0 \quad \text{as } k\longrightarrow\infty$$

Therefore

$$\lim_{k\to\infty}\|A_k\|=\|A\|$$

Corollary. The sequence $A_k\longrightarrow 0$ as $k\longrightarrow\infty$ if and only if

$$\lim_{k\to\infty}\|A_k\|=0$$

where $\|A_k\|$ is some canonical norm.

It is easy to see that if

$$\lim_{k\to\infty}A_k=A \quad \text{and} \quad \lim_{k\to\infty}B_k=B$$

then

(a) $\lim\limits_{k\to\infty}(A_k\pm B_k)=A\pm B$,

(b) $\lim\limits_{k\to\infty}(A_kB_k)=AB$,

(c) $\lim\limits_{k\to\infty}A_k^{-1}=A^{-1}$ $(\det A\neq 0)$

on the assumption that the corresponding operations are meaningful. In particular, if C is a constant matrix such that the multiplications CA_k and A_kC $(k=1, 2, \ldots)$ are possible, then

$$\lim_{k\to\infty}CA_k=CA$$

and

$$\lim_{k\to\infty}A_kC=AC$$

Lemma 2. *For the convergence of a sequence of matrices* A_k $(k = 1, 2, \ldots)$ *it is necessary and sufficient that the generalized Cauchy test hold*, namely: for any $\varepsilon > 0$ there must be a number $N = N(\varepsilon)$ such that for $k > N$, $p > 0$

$$\| A_{k+p} - A_k \| < \varepsilon \tag{5}$$

where $\| \quad \|$ is any canonical norm.

Indeed, if inequality (5) is valid, then for every element $a_{ij}^{(k)}$ of matrix A_k the Cauchy test (see Sec. 3.4) will hold, and hence there exists

$$\lim_{k \to \infty} A_k = [\lim_{k \to \infty} a_{ij}^{(k)}]$$

Conversely, if there exists

$$A = \lim_{k \to \infty} A_k$$

then by Lemma 1

$$\| A - A_k \| \to 0 \quad \text{as} \quad k \to \infty$$

and, thus, the inequality (5) holds true.

7.10 SERIES OF MATRICES

Using the concept of a limit of a matrix, we can introduce *series of matrices* (*matrix series*):

$$\sum_{k=1}^{\infty} A_k = \lim_{N \to \infty} \sum_{k=1}^{N} A_k \tag{1}$$

where A_k are matrices of the same dimensions.

If the limit (1) exists, then the matrix series is *convergent*, and the matrix obtained in the limit is termed the *sum of the series*. If the limit (1) does not exist, the matrix series is *divergent* and no sum is assigned to it.

A necessary condition for the convergence of a matrix series.

Theorem 1. *If the matrix series* (1) *converges, then*

$$\lim_{k \to \infty} A_k = 0$$

Proof. Let

$$S_k = \sum_{j=1}^{k} A_j$$

If the series (1) converges, then there exists the finite limit

$$S = \lim_{k \to \infty} S_k$$

We have

$$A_k = S_k - S_{k-1}$$

whence

$$\lim_{k \to \infty} A_k = \lim_{k \to \infty} S_k - \lim_{k \to \infty} S_{k-1} = S - S = 0$$

The matrix series (1) is called *absolutely convergent* if the following series converges:

$$\sum_{k=1}^{\infty} |A_k| \tag{2}$$

Theorem 2. *An absolutely convergent matrix series is a convergent series.*

Proof. Let

$$A_k = [a_{ij}^{(k)}] \qquad (k = 1, 2, \dots)$$

Then

$$\sum_{k=1}^{\infty} |A_k| = \left[\sum_{k=1}^{\infty} |a_{ij}^{(k)}| \right]$$

Since the matrix series (2) converges, then, by definition, each of the numerical series

$$\sum_{k=1}^{\infty} |a_{ij}^{(k)}| \; (i = 1, 2, \dots, m; \quad j = 1, 2, \dots. \; n)$$

is convergent. From this it follows by a familiar theorem of the theory of series that all the series $\sum_{k=1}^{\infty} a_{ij} \, (i = 1, \dots, m; j = 1, \dots, n)$ converge, and converge absolutely; that is, there is a limit

$$S = \lim_{N \to \infty} S_N = \lim_{N \to \infty} \sum_{k=1}^{N} A_k$$

and hence, the matrix series (1) converges.

For a rough analysis of the convergence of the matrix series (1) we can take advantage of the sufficient condition given below.

Theorem 3. *If* $\|A\|$ *is any canonical norm and the numerical series*

$$\sum_{k=1}^{\infty} \|A_k\| \tag{3}$$

converges, then the matrix series (1) *also converges, and converges absolutely.*

Proof. Let

$$A_k = [a_{ij}^{(k)}] \qquad (k = 1, 2, \dots)$$

We consider the numerical series

$$\sum_{k=1}^{\infty} a_{ij}^{(k)} \tag{4}$$

$(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$. Since

$$|a_{ij}^{(k)}| \leqslant \| A_k \|$$

it follows that each of the series (4) converges, and converges absolutely. Hence, the matrix series

$$\sum_{k=1}^{\infty} A_k = \left[\sum_{k=1}^{\infty} a_{ij}^{(k)}\right]$$

by definition converges, and converges absolutely.

Very important in applications are *matrix power series:* **right-hand,**

$$\sum_{k=0}^{\infty} A_k X^k \tag{5}$$

and **left-hand,**

$$\sum_{k=0}^{\infty} X^k A_k \tag{5'}$$

where X is a square matrix of order n. In the first case, A_k are $m \times n$ matrices or scalars (for instance, A_k may be row vectors); in the second case, A_k are $n \times m$ matrices or scalars (for instance, A_k may be column vectors).

Theorem 4. *If* r *is the radius of convergence of the scalar power series*

$$\sum_{k=0}^{\infty} \| A_k \| x^k \tag{6}$$

where $\| A_k \|$ $(k=0, 1, 2, \ldots)$ *is some canonical norm, then the matrix power series* (5) *and* (5′) *definitely converge for*

$$\| X \| < r \tag{7}$$

In particular, the matrix power series

$$\sum_{k=0}^{\infty} a_k X^k$$

with numerical coefficients a_k $(k=0, 1, 2, \ldots)$ converges for

$$\| X \| < r$$

where r is the radius of convergence of the power series

$$\sum_{k=0}^{\infty} |a_k| x^k$$

Proof. Since

$$\| A_k X^k \| \leqslant \| A_k \| \| X \|^k$$

then, when inequality (7) holds, the series

$$\sum_{k=0}^{\infty} \| A_k X^k \|$$

converges. Consequently, by Theorem 3, the power series (5) also converges.

Similar reasoning holds true for the series (5′).

The second assertion of the theorem follows from the fact that if a_k is a scalar, then

$$\| a_k \| = | a_k |$$

Theorem 5. *The geometric series*

$$A + AX + AX^2 + \ldots + AX^k + \ldots \tag{8}$$

and

$$A + XA + X^2A + \ldots + X^kA + \ldots \tag{8'}$$

where X *is a square matrix, converge if*

$$\| X \| < 1 \tag{9}$$

Here

$$\sum_{k=0}^{\infty} AX^k = A(E-X)^{-1}$$

and

$$\sum_{k=0}^{\infty} X^kA = (E-X)^{-1} A$$

Indeed, by Theorem 4, given the condition (9), the geometric series (8) converges, that is, there exists a finite matrix

$$S = \sum_{k=0}^{\infty} AX^k$$

Consider the identity

$$A(E + X + X^2 + \ldots + X^k)(E-X) = A(E - X^{k+1}) \tag{10}$$

Passing to the limit as $k \to \infty$ in (10) and noting that by virtue of Condition (9),

$$X^{k+1} \longrightarrow 0 \quad \text{as} \quad k \longrightarrow \infty$$

we will have

$$S(E-X)=AE=A \tag{11}$$

In particular, assuming $A=E$ in (11), we get

$$S_1(E-X)=E$$

where

$$S_1=\sum_{k=0}^{\infty} X^k$$

From this

$$\det S_1 \cdot \det(E-X)=\det E=1$$

Since $\det S_1$ is finite, it follows that

$$\det(E-X)\neq 0$$

and, consequently, the matrix $E-X$ is nonsingular, which means it has an inverse, $(E-X)^{-1}$.

Multiplying both members of (11) on the right by $(E-X)^{-1}$, we finally get

$$S=\sum_{k=0}^{\infty} AX^k=A(E-X)^{-1}$$

In a similar manner we prove that

$$\sum_{k=0}^{\infty} X^k A=(E-X)^{-1}A$$

for

$$\|X\|<1$$

Corollary. If $\|X\|<1$, then there is an inverse matrix

$$(E-X)^{-1}=\sum_{k=0}^{\infty} X^k$$

What is more, if $\|E\|=1$, then

$$\|(E-X)^{-1}\|\leqslant\sum_{k=0}^{\infty}\|X\|^k=\frac{1}{1-\|X\|}$$

Note. If $\|X\|<1$, then it is easy to estimate the norm of the remainder of the matrix series (8).

We have

$$R_k\equiv\|A(E-X)^{-1}-A(E+X+X^2+\ldots+X^k)\|\leqslant\|A\|\,\|X^{k+1}+$$
$$+X^{k+2}+\ldots\|\leqslant\|A\|(\|X\|^{k+1}+\|X\|^{k+2}+\ldots)=\frac{\|A\|\,\|X\|^{k+1}}{1-\|X\|}$$

Similarly for the series (8′) we have

$$R'_k=\|(E-X)^{-1}A-(E+X+X^2+\dots+X^k)A\|\leqslant\frac{\|A\|\,\|X\|^{k+1}}{1-\|X\|}$$

Matrix series make it possible to determine the *transcendental functions of a matrix*. For example, it is assumed that

$$e^X=\sum_{n=0}^{\infty}\frac{X^n}{n!} \tag{12}$$

and it is possible to prove that the series (12) converges for an arbitrary square matrix X.

7.11 Partitioned matrices

Suppose we have a matrix A. We partition it into matrices of lower orders (*submatrices*, or *blocks*) using horizontal and vertical partitions that run through the whole matrix. For example,

$$A=\left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ \hline a_{31} & a_{32} & a_{33}\end{array}\right]$$

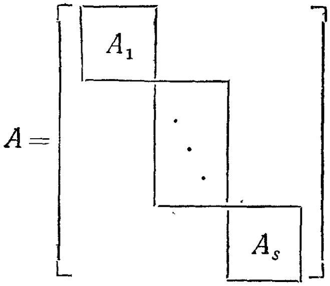
where the blocks are the submatrices

$$P=\begin{bmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix},\quad Q=\begin{bmatrix} a_{13}\\ a_{23}\end{bmatrix},\quad R=[a_{31}\ \ a_{32}],\quad S=[a_{33}]$$

Then A may be regarded as a supermatrix whose elements are blocks (submatrices):

$$A=\begin{bmatrix} P & Q\\ R & S\end{bmatrix}$$

A matrix partitioned into submatrices, or blocks, is called a *partitioned matrix*. Quite naturally, the partitioning of a matrix may be done in a variety of ways. A special case of partitioned matrices are the *quasidiagonal matrices*



where the blocks A_i $(i=1, \ldots, s)$ are square matrices of (generally speaking) different orders, all other elements being zeros. Note that

$$\det A = \det A_1 \ldots \det A_s$$

Another important special case of partitioned matrices are *bordered matrices*

$$A_n = \left[\begin{array}{c|c} A_{n-1} & U_n \\ \hline V_n & a_{nn} \end{array}\right]$$

where

$$A_{n-1} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,\,n-1} \\ a_{21} & a_{22} & \ldots & a_{2,\,n-1} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n-1,\,1} & a_{n-1,\,2} & \ldots & a_{n-1,\,n-1} \end{bmatrix}$$

is a matrix of order $n-1$;

$$U_n = \begin{bmatrix} a_{1,\,n} \\ a_{2,\,n} \\ \cdots \\ a_{n-1,\,n} \end{bmatrix} \quad \text{is a column matrix;}$$

$V_n = [a_{n,\,1}\ a_{n,\,2} \ldots a_{n,\,n-1}]$ is a row matrix and a_{nn} is a scalar.

Let us agree to use the term *conformal* to designate partitioned matrices of the same dimensions and with the same partitioning. Partitioned matrices are convenient in that operations involving them are carried out formally by the same rules used for ordinary matrices.

A. THE ADDITION AND SUBTRACTION OF PARTITIONED MATRICES

If the partitioned matrices

$$A = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1q} \\ \cdot & \cdot & \cdot & \cdot \\ A_{p1} & A_{p2} & \ldots & A_{pq} \end{bmatrix} \tag{1}$$

and

$$B = \begin{bmatrix} B_{11} & B_{12} & \ldots & B_{1s} \\ \cdot & \cdot & \cdot & \cdot \\ B_{r1} & B_{r2} & \ldots & B_{rs} \end{bmatrix} \tag{2}$$

are conformal, that is, $p=r$, $q=s$ and the blocks A_{ij} and B_{ij} have the same dimensions, then

$$A+B=\begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} & \dots & A_{1q}+B_{1q} \\ \cdot & \cdot & \cdot & \cdot \\ A_{p1}+B_{p1} & A_{p2}+B_{p2} & \dots & A_{pq}+B_{pq} \end{bmatrix}$$

Indeed, in order to add the matrices A and B it is necessary to add the corresponding elements, but it is obvious that the same thing is achieved if we add the corresponding blocks (submatrices) of these matrices.

Subtraction of partitioned matrices is performed analogously.

If A is the partitioned matrix (1) and α is a scalar, then we have

$$\alpha A=\begin{bmatrix} \alpha A_{11} & \alpha A_{12} & \dots & \alpha A_{1q} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha A_{p1} & \alpha A_{p2} & \dots & \alpha A_{pq} \end{bmatrix}$$

B. THE MULTIPLICATION OF PARTITIONED MATRICES

Suppose the partitioned matrices A and B have the structure given in (1) and (2), respectively, and $q=r$.

Assume that all the blocks A_{ij} and B_{jk} $(i=1, 2, \dots, p;$ $j=1, 2, \dots, q;$ $k=1, 2, \dots, s)$ are such that the number of columns of block A_{ij} is equal to the number of rows of block B_{jk}. In the special case when all blocks A_{ij} and B_{ij} are square and have the same order, this assumption is definitely fulfilled. Then we can prove that the product of the matrices A and B is the partitioned matrix

$$C=\begin{bmatrix} C_{11} & C_{12} & \dots & C_{1s} \\ C_{21} & C_{22} & \dots & C_{2s} \\ \cdot & \cdot & \cdot & \cdot \\ C_{p1} & C_{p2} & \dots & C_{ps} \end{bmatrix}$$

where $C_{ik}=A_{i1}B_{1k}+A_{i2}B_{2k}+\dots+A_{iq}B_{qk}$ $(i=1, 2, \dots, p;$ $k=1, 2, \dots, s)$, that is, the matrices A and B are multiplied together as if the blocks (submatrices) were numbers [2].

Example. Multiply the partitioned matrices

$$A=\left[\begin{array}{c|c|c} & \leftarrow 2 \rightarrow & \leftarrow 1 \rightarrow \\ \hline \updownarrow 2 & P & Q \end{array}\right]$$

and

$$B=\left[\begin{array}{c|c|c} & \leftarrow 1\rightarrow & \leftarrow 2\rightarrow \\ \hline \overset{\uparrow}{\underset{\downarrow}{2}} & R & S \\ \hline \overset{\uparrow}{\underset{\downarrow}{1}} & T & U \end{array}\right]$$

to get a matrix of the form

$$AB=\left[\begin{array}{c|c|c} & \leftarrow 1\rightarrow & \leftarrow 2\rightarrow \\ \hline \overset{\uparrow}{\underset{\downarrow}{2}} & PR+QT & PS+QU \end{array}\right]$$

Addition and multiplication of quasidiagonal matrices are especially simple. If

$$A=\begin{bmatrix} \boxed{A_1} & & \\ & \ddots & \\ & & \boxed{A_s} \end{bmatrix}, \qquad B=\begin{bmatrix} \boxed{B_1} & & \\ & \ddots & \\ & & \boxed{B_s} \end{bmatrix}$$

and the orders of the matrices A_i, B_i $(i=1, 2, \ldots, s)$ are the same, then we clearly have

$$A+B=\begin{bmatrix} \boxed{A_1+B_1} & & \\ & \ddots & \\ & & \boxed{A_s+B_s} \end{bmatrix}$$

and

$$AB=\begin{bmatrix} \boxed{A_1B_1} & & \\ & \boxed{\ddots} & \\ & & \boxed{A_sB_s} \end{bmatrix}$$

7.12 MATRIX INVERSION BY PARTITIONING

Suppose for a given nonsingular numerical matrix A it is required to find the inverse A^{-1}. Partition A into four submatrices:

$$A=\begin{bmatrix} \alpha_{11}(r,\ r) & \alpha_{12}(r,\ s) \\ \alpha_{21}(s,\ r) & \alpha_{22}(s,\ s) \end{bmatrix}$$

The orders of the submatrices are indicated in parentheses; and $r+s=n$, where n is the order of matrix A. We seek the inverse A^{-1} also in the form of a four-block matrix,

$$A^{-1}=\begin{bmatrix} \beta_{11}(r,\ r) & \beta_{12}(r,\ s) \\ \beta_{21}(s,\ r) & \beta_{22}(s,\ s) \end{bmatrix}$$

Then, since $A^{-1}A=E$, by multiplying the matrices we get four matrix equations:

$$\left.\begin{aligned} \beta_{11}\alpha_{11}+\beta_{12}\alpha_{21}&=E_r, \\ \beta_{11}\alpha_{12}+\beta_{12}\alpha_{22}&=0, \\ \beta_{21}\alpha_{11}+\beta_{22}\alpha_{21}&=0, \\ \beta_{21}\alpha_{12}+\beta_{22}\alpha_{22}&=E_s \end{aligned}\right\} \tag{1}$$

where E_r and E_s are unit matrices of appropriate orders. Solving this system, we determine the blocks of matrix A^{-1}. In solving (1), we use the method of eliminating unknowns. Postmultiplying the first equation of (1) by $\alpha_{11}^{-1}\alpha_{12}$ and subtracting the second equation of the system from the result obtained, we have

$$\beta_{12}(\alpha_{21}\alpha_{11}^{-1}\alpha_{12}-\alpha_{22})=\alpha_{11}^{-1}\alpha_{12}$$

whence we find

$$\beta_{12}=-\alpha_{11}^{-1}\alpha_{12}(\alpha_{22}-\alpha_{21}\alpha_{11}^{-1}\alpha_{12})^{-1}$$

and

$$\beta_{11}=\alpha_{11}^{-1}-\beta_{12}\alpha_{21}\alpha_{11}^{-1}$$

Similarly, from the third and fourth equations of (1) we get

$$\beta_{22} = (\alpha_{22} - \alpha_{21}\alpha_{11}^{-1}\alpha_{12})^{-1}$$

and

$$\beta_{21} = -\beta_{22}\alpha_{21}\alpha_{11}^{-1}$$

It is of course assumed here that the corresponding operations are meaningful. We introduce the matrices

$$\left.\begin{array}{ll} X = \alpha_{11}^{-1}\alpha_{12}, & Y = \alpha_{21}\alpha_{11}^{-1}, \\ \theta = \alpha_{22} - \alpha_{21} & X = \alpha_{22} - Y\alpha_{12} \end{array}\right\} \qquad (2)$$

Then the formulas for the blocks β_{ij} $(i, j = 1, 2)$ may be written more simply:

$$\begin{aligned} &\beta_{11} = \alpha_{11}^{-1} + X\theta^{-1}Y, \\ &\beta_{12} = -X\theta^{-1}, \\ &\beta_{21} = -\theta^{-1}Y, \ \beta_{22} = \theta^{-1} \end{aligned}$$

Formulas (1) determine the blocks of matrix A^{-1} provided that α_{11}^{-1} and θ^{-1} exist. It is convenient to arrange the computations in the following scheme [4]:

	α_{21}	α_{22}
$X = \alpha_{11}^{-1}\alpha_{12}$	α_{11}^{-1}	α_{12}
θ^{-1}	$Y = \alpha_{21}\alpha_{11}^{-1}$	$\theta = \alpha_{22} - Y\alpha_{12}$

and

$$A^{-1} = \left[\begin{array}{c|c} \alpha_{11}^{-1} + X\theta^{-1}Y & -X\theta^{-1} \\ \hline -\theta^{-1}Y & \theta^{-1} \end{array}\right]$$

This method is useful if matrix α_{11} is readily invertible.

Example 1. Invert the matrix

$$\begin{bmatrix} 1 & 0 & 3 & -4 \\ 0 & 1 & 5 & 6 \\ -3 & 4 & 0 & 2 \\ -5 & -6 & 2 & 0 \end{bmatrix}$$

Solution. Set

$$\alpha_{11}=\begin{bmatrix}1&0\\0&1\end{bmatrix},\qquad \alpha_{12}=\begin{bmatrix}3&-4\\5&6\end{bmatrix},$$
$$\alpha_{21}=\begin{bmatrix}-3&4\\-5&-6\end{bmatrix},\quad \alpha_{22}=\begin{bmatrix}0&2\\2&0\end{bmatrix}$$

Using the above scheme, we have

$$\begin{array}{c|cc|cc|cc|}
 & & & -3 & 4 & 0 & 2\\
 & & & -5 & -6 & 2 & 0\\
\hline
X & 3 & -4 & 1 & 0 & 3 & -4\\
 & 5 & 6 & 0 & 1 & 5 & 6\\
\hline
\theta^{-1}\frac{1}{1422} & 16 & 34 & -3 & 4 & -11 & -34\\
 & -47 & -11 & -5 & -6 & 47 & 16\\
\hline
\multicolumn{1}{c}{} & & \multicolumn{1}{c}{} & \multicolumn{2}{c}{Y} & \multicolumn{2}{c}{\theta}
\end{array}$$

Whence

$$X\theta^{-1}=\frac{1}{1422}\begin{bmatrix}3&-4\\5&6\end{bmatrix}\begin{bmatrix}16&34\\-47&-11\end{bmatrix}=\frac{1}{1422}\begin{bmatrix}236&146\\-202&104\end{bmatrix},$$
$$\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix}16&34\\-47&-11\end{bmatrix}\begin{bmatrix}-3&4\\-5&-6\end{bmatrix}=\frac{1}{1422}\begin{bmatrix}-218&-140\\196&-122\end{bmatrix},$$
$$X\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix}236&146\\-202&104\end{bmatrix}\begin{bmatrix}-3&4\\-5&-6\end{bmatrix}=\frac{1}{1422}\begin{bmatrix}-1438&68\\86&-1432\end{bmatrix}$$

As a check, we compute the product $X\theta^{-1}Y$ in two ways:

$$X\theta^{-1}Y=(X\theta^{-1})Y \quad\text{and}\quad Y\theta^{-1}Y=X(\theta^{-1}Y)$$

By the general scheme we have

$$A^{-1}=\frac{1}{1422}\left[\begin{array}{cc|cc}-16&68&-236&-146\\86&-10&202&-104\\\hline 218&140&16&34\\-196&122&-47&-11\end{array}\right]$$

A particular case of the foregoing method is the so-called *method of bordering*. Essentially it is this. Suppose we have a matrix

$$A=\begin{bmatrix}a_{11}\dots a_{1n}\\ \cdots\cdots\\ a_{n1}\dots a_{nn}\end{bmatrix}$$

We form the following sequence of matrices:

$$S_1=[a_{11}],$$

$$S_2=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

$$S_3=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\left[\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{S_2}} & a_{13} \\ & & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right],$$

$$S_4=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}=\left[\begin{array}{ccc|c} & & & a_{14} \\ & S_3 & & a_{24} \\ & & & a_{34} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right]$$

and so on. Each matrix is obtained from the preceding one by bordering. The inverse of the second of these matrices, S^{-1}, is found directly:

$$S_2^{-1}=\begin{bmatrix} \dfrac{a_{22}}{\Delta} & -\dfrac{a_{12}}{\Delta} \\ -\dfrac{a_{21}}{\Delta} & \dfrac{a_{11}}{\Delta} \end{bmatrix}$$

where

$$\Delta=a_{11}a_{22}-a_{12}a_{21}$$

Then, applying S_2^{-1} to S_3 by the foregoing scheme of computation, we get S_3^{-1}; then we use S_3^{-1} to get S_4^{-1}, and, finally, $S_n^{-1}=A^{-1}$.

The method of bordering cannot be applied if one of the intermediate matrices S_i is singular. The situation can, however, be rectified by an interchange of the rows of the matrix [5].

Example 2. Find the inverse of the matrix

$$A=\begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{bmatrix}$$

Solution. Here

$$S_2=\begin{bmatrix} 1 & 4 \\ 0 & -1 \end{bmatrix}=S_2^{-1}$$

The scheme for computing S_3^{-1} is as follows:

$$\begin{array}{c|r|rr|r|}
\cline{3-5}
 & & 3 & 1 & 0 \\
\cline{2-5}
X & 13 & 1 & 4 & 1 \\
 & -3 & 0 & -1 & 3 \\
\cline{2-5}
\theta^{-1} & -\frac{1}{36} & 3 & 11 & -36 \\
\cline{2-5}
\multicolumn{2}{c}{} & \multicolumn{2}{c}{Y} & \multicolumn{1}{c}{\theta}
\end{array}$$

$$X\theta^{-1}Y=\begin{bmatrix} -\frac{13}{12} & -\frac{143}{36} \\ \frac{1}{4} & \frac{11}{12} \end{bmatrix}$$

Hence

$$S_3^{-1}=\left[\begin{array}{rr|r} -\frac{1}{12} & \frac{1}{36} & \frac{13}{36} \\ \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} \\ \hline \frac{1}{12} & \frac{11}{36} & -\frac{1}{36} \end{array}\right]$$

The following scheme can be used to compute S_4^{-1}:

$$\begin{array}{c|r|rrr|r|}
\cline{3-6}
 & & 1 & -2 & 5 & 1 \\
\cline{2-6}
 & \frac{4}{9} & -\frac{1}{12} & \frac{1}{36} & \frac{13}{36} & 3 \\
X & \frac{2}{3} & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -1 \\
 & -\frac{1}{9} & \frac{1}{12} & \frac{11}{36} & \frac{1}{36} & 2 \\
\cline{2-6}
\theta^{-1} & \frac{9}{22} & -\frac{1}{6} & \frac{31}{18} & \frac{7}{18} & \frac{22}{9} \\
\cline{2-6}
\multicolumn{2}{c}{} & \multicolumn{3}{c}{Y} & \multicolumn{1}{c}{\theta}
\end{array}$$

$$X\theta^{-1}Y=\begin{bmatrix} -\frac{1}{33} & \frac{31}{99} & \frac{7}{99} \\ -\frac{1}{22} & \frac{31}{66} & \frac{7}{66} \\ \frac{1}{132} & -\frac{31}{396} & -\frac{7}{396} \end{bmatrix}$$

Hence

$$S_4^{-1}=A^{-1}=\left[\begin{array}{ccc|c} -\frac{5}{44} & \frac{15}{44} & \frac{19}{44} & -\frac{2}{11} \\ \frac{9}{44} & \frac{17}{44} & \frac{1}{44} & -\frac{3}{11} \\ \frac{4}{44} & \frac{10}{44} & -\frac{2}{44} & \frac{1}{22} \\ \hline \frac{3}{44} & -\frac{31}{44} & -\frac{7}{44} & \frac{9}{22} \end{array}\right]=$$

$$=\frac{1}{44}\begin{bmatrix} -5 & 15 & 19 & -8 \\ 9 & 17 & 1 & -12 \\ 4 & 10 & -2 & 2 \\ 3 & -31 & -7 & 18 \end{bmatrix}$$

7.13 TRIANGULAR MATRICES

Definition. A square matrix is called *triangular* if the elements above (below) the main diagonal are zero. For example,

$$T=\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & t_{nn} \end{bmatrix}$$

where $t_{ij}=0$ for $i>j$ is an upper triangular matrix. Similarly,

$$T_1=\begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{21} & t_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{bmatrix}$$

where $t_{ij}=0$ for $j>i$ is a lower triangular matrix.

A diagonal matrix is a particular case of both an upper and lower triangular matrix. The determinant of a triangular matrix is equal to the product of its diagonal elements, namely, if $T=[t_{ij}]$ is a triangular matrix, then we obviously have $\det T=t_{11}t_{22}\dots t_{nn}$. Therefore, a triangular matrix is nonsingular only when all its diagonal elements are nonzero.

It can be proved that (1) the sum and product of triangular matrices of the same dimensions and the same structure (that is, upper only or lower only) are also triangular matrices of the same dimensions and general structure; (2) the inverse of a nonsingular triangular matrix is also a triangular matrix of the same dimensions and structure. Utilizing this circumstance, we can readily invert a triangular matrix.

Example 1. Invert the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

Solution. Set

$$A^{-1} = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix}$$

Multiplying the matrices A and A^{-1}, we get

$$\left.\begin{array}{rr} t_{11} = 1, & t_{11} + 2t_{21} + 3t_{31} = 0, \\ t_{11} + 2t_{21} = 0, & 2t_{22} + 3t_{32} = 0, \\ 2t_{22} = 1 & 3t_{33} = 1 \end{array}\right\}$$

whence we successively find

$$t_{11} = 1, \quad t_{21} = -\frac{1}{2}, \quad t_{22} = \frac{1}{2},$$
$$t_{31} = 0, \quad t_{32} = -\frac{1}{3}, \quad t_{33} = \frac{1}{3}$$

Consequently

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

The following important theorem holds true [3].

Theorem. *Any square matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

with nonzero principal-diagonal minors

$$\Delta_1 = a_{11} \neq 0, \ \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \ \dots, \ \Delta_n = |A| \neq 0$$

may be represented as a product of two triangular matrices of diffe-

rent structure (lower and upper); this expansion will be unique if the diagonal elements of one of the triangular matrices are fixed beforehand (say by putting them equal to 1).

We omit the proof but indicate a method for finding the elements of the desired triangular matrices. Let

$$A = T_1 T_2 \tag{1}$$

where

$$T_1 = [b_{ij}], \quad b_{ij} = 0 \text{ when } j > i \tag{2}$$

is a lower triangular matrix of order n;

$$T_2 = [c_{ij}], \quad c_{ij} = 0 \quad \text{when } i > j \tag{3}$$

is an upper triangular matrix of order n. Multiplying together these matrices we get, by formula (1),

$$\sum_{k=1}^{n} b_{ik} c_{kj} = a_{ij} \qquad (i, j = 1, 2, \ldots, n) \tag{4}$$

Due to Conditions (2) and (3), system (4) becomes

$$\sum_{k=1}^{j} b_{ik} c_{kj} = a_{ij} \quad \text{for} \quad i \geqslant j \qquad (j = 1, 2, \ldots, n) \tag{4'}$$

and

$$\sum_{k=1}^{i} b_{ik} c_{kj} = a_{ij} \quad \text{for} \quad i < j \qquad (i = 1, 2, \ldots, n-1) \tag{4''}$$

Because of their peculiar structure, the systems (4′) and (4″) are readily solved to within the diagonal elements b_{ii} and c_{ii}. For the sake of definiteness, we can put $c_{ii} = 1$ $(i = 1, 2, \ldots, n)$.

Example 2. Represent the matrix

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix}$$

as a product of two triangular matrices T_1 and T_2.

Solution. $A = T_1 T_2$. We seek T_1 and T_2 in the form

$$T_1 = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \quad \text{and} \quad T_2 = \begin{bmatrix} 1 & r_{12} & r_{13} \\ 0 & 1 & r_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

We have

$$\begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{11} r_{12} & t_{11} r_{13} \\ t_{21} & t_{21} r_{12} + t_{22} & t_{21} r_{13} + t_{22} r_{23} \\ t_{31} & t_{31} r_{12} + t_{32} & t_{31} r_{13} + t_{32} r_{23} + t_{33} \end{bmatrix}$$

whence

$$\begin{array}{lll} t_{11}=1, & t_{11}r_{12}=-1, & t_{11}r_{13}=2, \\ t_{21}=-1, & t_{21}r_{12}+t_{22}=5, & t_{21}r_{13}+t_{22}r_{23}=4, \\ t_{31}=2, & t_{31}r_{12}+t_{32}=4, & t_{31}r_{13}+t_{32}r_{23}+t_{33}=14 \end{array}$$

Solving the system, we get

$$\begin{array}{lll} t_{11}=1, & t_{21}=-1, & t_{31}=2, \\ t_{22}=4, & t_{32}=6, & t_{33}=1, \\ r_{12}=-1, & r_{13}=2, & r_{23}=\frac{3}{2} \end{array}$$

Thus

$$T_1=\begin{bmatrix} 1 & 0 & 0 \\ -1 & 4 & 0 \\ 2 & 6 & 1 \end{bmatrix}$$

and

$$T_2=\begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1 \end{bmatrix}$$

Using the representation of a square matrix A (det $A \neq 0$) as a product of two triangular matrices, we can indicate another method for computing the inverse A^{-1}; namely, if

$$A=T_1T_2$$

then

$$A^{-1}=T_2^{-1}T_1^{1}$$

We thus see that finding the inverses of triangular matrices is a comparatively simple affair.

7.14 ELEMENTARY TRANSFORMATIONS OF MATRICES

The following matrix transformations are called *elementary:*

(1) interchanging two rows or columns;

(2) multiplying all elements of one row (column) by the same nonzero number (scalar);

(3) adding multiples of the elements of a row (column) to the elements of another row (column).

Two matrices are termed *equivalent* if one is obtained from the other via a finite number of elementary transformations. Such matrices are not, generally speaking, equal, but, as may be proved, they have the same rank [6].

It is easy to see that every elementary transformation of a square matrix A is equivalent to multiplying A by some nonsingular matrix. Then, if the transformation operation is performed on rows (columns) of A, the multiplier must be a postmultiplier (premultiplier) and be the result of applying the corresponding elementary transformation to the unit matrix [6]. For example, if in the matrix

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

we interchange the second and third rows, we get the equivalent matrix

$$\tilde{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

The same matrix $\tilde{A}$ is obtained if we interchange the second and third rows in the unit matrix

$$E=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and postmultiply the resulting matrix

$$\tilde{E}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

by the matrix A, that is $\tilde{A}=\tilde{E}A$.

Other elementary transformations are carried out in a similar manner. Note that if in the equation $AA^{-1}=E$ we perform identical transformations of the rows of the matrices A and E until A becomes a unit matrix, then we will have $\tilde{E}AA^{-1}=\tilde{E}$, where $\tilde{E}$ is the transformed unit matrix. Then, since $\tilde{E}A=E$, we get $A^{-1}=\tilde{E}$, that is to say, the inverse matrix A^{-1} is a transformed unit matrix. This is the basis of the method of computing an inverse by means of row transformations [4].

7.15 COMPUTATION OF DETERMINANTS

Elementary matrix transformations offer the most convenient method for computing the determinant of a matrix. Suppose, for example, we have

$$\Delta_n=\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} \tag{1}$$

Assuming that $a_{11} \neq 0$, we have

$$\Delta_n = a_{11} \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ \frac{a_{21}}{a_{11}} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{a_{n1}}{a_{11}} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Here, subtracting from the elements a_{ij} of the jth column $(j \geqslant 2)$ the corresponding elements of the first column multiplied by a_{1j}, we get

$$\Delta_n = a_{11} \begin{bmatrix} 1 & 0 & \dots & 0 \\ \frac{a_{21}}{a_{11}} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{a_{n1}}{a_{11}} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix} = a_{11}\Delta_{n-1}$$

where

$$\Delta_{n-1} = \begin{vmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} \tag{2}$$

and

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \qquad (i,\ j = 2,\ 3,\ \dots,\ n)$$

Apply the same technique to the determinant Δ_{n-1}. If all the elements

$$a_{ii}^{(i-1)} \neq 0 \qquad (i = 1,\ 2,\ \dots,\ n)$$

we finally obtain

$$\Delta_n = a_{11}a_{22}^{(1)} \dots a_{nn}^{(n-1)} \tag{3}$$

If in some intermediate determinant Δ_{n-k} the upper left element $a_{k+1,\,k+1}^{(k)} = 0$, then it is necessary to interchange the rows or columns of the determinant Δ_{n-k} so that the element we need is nonzero (this is always possible if the determinant $\Delta \neq 0$). Of course we must take into consideration the change in the sign of Δ_{n-k}. We can give a more general rule. Suppose the determinant $\tilde{\Delta}_n = \det[\alpha_{ij}]$ is transformed so that $\alpha_{pq} = 1$ (α_{pq} is the principal element), that is,

$$\tilde{\Delta}_n = \begin{vmatrix} \alpha_{11} \dots \alpha_{1q} & \dots \alpha_{1j} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \alpha_{i1} \dots \boxed{\alpha_{iq}} & \dots \alpha_{ij} & \dots & \alpha_{in} \\ \dots & \dots & \dots & \dots \\ \alpha_{p1} \dots 1 & \dots \boxed{\alpha_{pj}} & \dots & \alpha_{pn} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} \dots \alpha_{nq} & \dots \alpha_{nj} & \dots & \alpha_{nn} \end{vmatrix}$$

Then

$$\tilde{\Delta}_n = (-1)^{p+q}\, \tilde{\Delta}_{n-1}$$

where $\tilde{\Delta}_{n-1} = \det [\alpha_{ij}^{(1)}]$ is a determinant of the $(n-1)$th order obtained from $\tilde{\Delta}_n$ by deleting the pth row and the qth column with a subsequent transformation of the elements via the formula

$$\alpha_{ij}^{(1)} = \alpha_{ij} - \alpha_{iq}\alpha_{pj}$$

Thus, *each element* $\alpha_{ij}^{(1)}$ *of the determinant* $\tilde{\Delta}_{n-1}$ *is equal to the corresponding element* α_{ij} *of the determinant* $\tilde{\Delta}_n$ *diminished by the product of its "projections"* α_{iq} and α_{pj} *by the discarded column and row of the original determinant*. The proof of this proposition follows readily from the general properties of determinants [7].

Example. Evaluate

$$\Delta_5 = \begin{vmatrix} 3 & 1 & -1 & 2 & \boxed{1} \\ -2 & 3 & 1 & 4 & 3 \\ 1 & 4 & 2 & 3 & 1 \\ 5 & -2 & -3 & 5 & -1 \\ -1 & 1 & 2 & 3 & 2 \end{vmatrix}$$

Solution. Taking $a_{15} = 1$ for the principal element, we have

$$\Delta_5 = (-1)^{1+5} \begin{vmatrix} -2-3\cdot 3 & 3-1\cdot 3 & 1-(-1)\cdot 3 & 4-2\cdot 3 \\ 1-3\cdot 1 & 4-1\cdot 1 & 2-(-1)\cdot 1 & 3-2\cdot 1 \\ 5-3\cdot(-1) & -2-1\cdot(-1) & -3-(-1)\cdot(-1) & 5-2\cdot(-1) \\ -1-3\cdot 2 & 1-1\cdot 2 & 2-(-1)\cdot 2 & 3-2\cdot 2 \end{vmatrix} =$$

$$= \begin{vmatrix} -11 & 0 & 4 & -2 \\ -2 & 3 & 3 & \boxed{1} \\ 8 & -1 & -4 & 7 \\ -7 & -1 & 4 & -1 \end{vmatrix}$$

Now taking $a_{24}=1$ for the principal element and applying a similar transformation, we obtain

$$\Delta_4=(-1)^6\begin{vmatrix} -15 & 6 & 10 \\ 22 & -22 & -25 \\ -9 & 2 & 7 \end{vmatrix}=2\begin{vmatrix} -15 & 3 & 10 \\ 22 & -11 & -25 \\ -9 & \boxed{1} & 7 \end{vmatrix}=$$

$$=2\cdot(-1)^{3+2}\begin{vmatrix} 12 & -11 \\ -77 & 52 \end{vmatrix}=446$$

Note that the number of multiplications and divisions required in the computation of an nth-order determinant is [8] equal to

$$\frac{n-1}{3}(n^2+n+3)$$

REFERENCES FOR CHAPTER 7

[1] *O. Schreier und E. Sperner, Vorlesungen über Matrizen,* 1932, Secs. 1, 2

[2] *A. Maltsev, Principles of Linear Algebra,* 1956 (in Russian).

[3] *V. N. Faddeyeva, Computational Methods of Linear Algebra,* 1950 (in Russian).

[4] *R. A. Frazer, W. J. Duncan, A. R. Collar, Elementary Matrices and Some Applications to Dynamics and Differential Equations,* 1946.

[5] *B. V. Bulgakov, Oscillations,* 1954, Chapter I (in Russian).

[6] *E. S. Lyapin, Course of Higher Algebra,* 1953, Chapter IX (in Russian).

[7] *E. T. Whittaker and G. Robinson, The Calculus of Observations,* 1944, Chapter V.

[8] *D. K. Faddeyev and V. N. Faddeyeva, Computational Methods of Linear Algebra,* 1960, Chapter II (in Russian).

Chapter 8

SOLVING SYSTEMS OF LINEAR EQUATIONS

8.1 A GENERAL DESCRIPTION OF METHODS OF SOLVING SYSTEMS OF LINEAR EQUATIONS

Methods of solving systems of linear equations are divided mainly into two groups: (1) *exact methods*, which are finite algorithms for computing the roots of a system (such, for example, are *Cramer's rule*, the *Gaussian method*, the *method of principal elements*, the *method of square roots*, etc.), and (2) *iterative methods*, which permit obtaining the roots of a system to a given accuracy by means of convergent infinite processes (these include the *method of iteration*, the *Seidel method*, the *method of relaxation*, and others).

Because of unavoidable rounding, the results of even the exact methods are **approximate**, and error estimates of the roots are, in the general case, involved. In the case of iterative processes we have the added error of method.

It will be noted that the effective use of iterative methods depends essentially on the apt choice of the initial approximation and the speed of convergence of the process.

8.2 SOLUTION BY INVERSION OF MATRICES. CRAMER'S RULE

Suppose we have a system of n linear equations in n unknowns:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{array}\right\} \tag{1}$$

Denote the matrix of the coefficients of (1) by

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \tag{2}$$

the column of its constant terms by

$$\boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \tag{3}$$

and the column of the unknowns (the **desired vector**) by

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{4}$$

Then the system (1) can be compactly written in the form of a matrix equation:

$$A\boldsymbol{x}=\boldsymbol{b} \tag{5}$$

The set of numbers $x_1, x_2, \ldots, x_n$ (or, briefly, the vector $\boldsymbol{x}$) which reduce (1) to an identity is called the *solution set* of the system, and the numbers x_i are termed the *roots* of the system.

If the matrix A is nonsingular, that is,

$$\det A=\begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}=\Delta\neq 0 \tag{6}$$

then (1), or the matrix equation (5) equivalent to it, has a unique solution.

Indeed, provided $\det A\neq 0$, there is an inverse matrix A^{-1}. Premultiplying both members of (5) by the matrix A^{-1}, we obtain

$$A^{-1}A\boldsymbol{x}=A^{-1}\boldsymbol{b}$$

or

$$\boldsymbol{x}=A^{-1}\boldsymbol{b} \tag{7}$$

Formula (7) plainly yields a solution of (5) and since every solution is of the form (7), the solution is unique.

Example 1. Solve the system of equations

$$\left.\begin{aligned} 3x_1-x_2&=5, \\ -2x_1+x_2+x_3&=0, \\ 2x_1-x_2+4x_3&=15 \end{aligned}\right\}$$

Solution. Write the system in matrix form:

$$\begin{bmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix}$$

The determinant of matrix A of the given system is

$$\det A = \begin{vmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{vmatrix} = 5 \neq 0$$

Computing the inverse matrix A^{-1}, we obtain

$$A^{-1} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix}$$

whence

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Thus, $x_1 = 2$, $x_2 = 1$, $x_3 = 3$.

A great deal of time is required to find the inverse A^{-1} of a matrix A of order $n > 4$ directly. For this reason, formula (7) is rarely used for practical purposes.

Using formula (7), it is easy to obtain formulas for the unknowns of the system (1). As we know (see Sec. 7.4),

$$A^{-1} = \frac{1}{\Delta} \tilde{A}$$

where

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ \cdot & \cdot & \cdot & \cdot \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}$$

is the adjoint of A (A_{ij} are the cofactors of the elements a_{ij}). Therefore

$$\boldsymbol{x} = \frac{1}{\Delta} \tilde{A} \boldsymbol{b}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \cdot \\ \cdot \\ \cdot \\ \Delta_n \end{bmatrix} \tag{8}$$

where

$$\Delta_i = \sum_{j=1}^{n} A_{ji} b_j = \begin{vmatrix} a_{11} & \dots & a_{1,\,i-1} & b_1 & a_{1,\,i+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,\,i-1} & b_2 & a_{2,\,i+1} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \dots & a_{n,\,i-1} & b_n & a_{n,\,i+1} & \dots & a_{nn} \end{vmatrix}$$

are the determinants obtained from the determinant Δ [formula (6)] by replacing its ith column by the column of constant terms of system (1). From equation (8) we get *Cramer's formulas*:

$$x_1 = \frac{\Delta_1}{\Delta}, \quad x_2 = \frac{\Delta_2}{\Delta}, \quad \dots, \quad x_n = \frac{\Delta_n}{\Delta} \tag{9}$$

Thus, if the determinant of system (1) is nonzero, $\Delta \neq 0$, then the system has a unique solution $\boldsymbol{x}$ defined by the matrix formula (7) or by the scalar formulas (9) equivalent to it.

Example 2. Solve the system of linear equations

$$\left.\begin{aligned} 2x_1 + x_2 - 5x_3 + x_4 &= 8, \\ x_1 - 3x_2 - 6x_4 &= 9, \\ 2x_2 - x_3 + 2x_4 &= -5, \\ x_1 + 4x_2 - 7x_3 + 6x_4 &= 0 \end{aligned}\right\}$$

Solution. The determinant of this system is

$$\Delta = \begin{vmatrix} 2 & 1 & -5 & 1 \\ 1 & -3 & 0 & -6 \\ 0 & 2 & -1 & 2 \\ 1 & 4 & -7 & 6 \end{vmatrix} = 27 \neq 0$$

Computing the supplementary determinants, we get

$$\Delta_1 = \begin{vmatrix} 8 & 1 & -5 & 1 \\ 9 & -3 & 0 & -6 \\ -5 & 2 & -1 & 2 \\ 0 & 4 & -7 & 6 \end{vmatrix} = 81,$$

$$\Delta_2 = \begin{vmatrix} 2 & 8 & -5 & 1 \\ 1 & 9 & 0 & -6 \\ 0 & -5 & -1 & 2 \\ 1 & 0 & -7 & 6 \end{vmatrix} = -108,$$

$$\Delta_3 = \begin{vmatrix} 2 & 1 & 8 & 1 \\ 1 & -3 & 9 & -6 \\ 0 & 2 & -5 & 2 \\ 1 & 4 & 0 & 6 \end{vmatrix} = -27,$$

$$\Delta_4 = \begin{vmatrix} 2 & 1 & -5 & 8 \\ 1 & -3 & 0 & 9 \\ 0 & 2 & -1 & -5 \\ 1 & 4 & -7 & 0 \end{vmatrix} = 27$$

whence

$$x_1 = \frac{\Delta_1}{\Delta} = \frac{81}{27} = 3,$$

$$x_2 = \frac{\Delta_2}{\Delta} = -\frac{108}{27} = -4,$$

$$x_3 = \frac{\Delta_3}{\Delta} = -\frac{27}{27} = -1,$$

$$x_4 = \frac{\Delta_4}{\Delta} = \frac{27}{27} = 1$$

Thus, the solution of a linear system (1) in n unknowns reduces to evaluating the $(n+1)$th determinant of order n. If n is great, evaluating the determinants is a laborious operation. For this reason, direct techniques have been elaborated for finding the roots of a linear system of equations.

8.3 THE GAUSSIAN METHOD

The most common technique for the solution of systems of linear equations is via an algorithm for the successive elimination of the unknowns. This method is called the *Gaussian method*. For the sake of simplicity, we confine ourselves to a system of four equations in four unknowns:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= a_{15}, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= a_{25}, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= a_{35}, \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= a_{45} \end{aligned}\right\} \quad (1)$$

Let the **leading element** $a_{11} \neq 0$. Dividing the coefficients of the first equation of (1) by a_{11}, we get

$$x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 = b_{15} \tag{2}$$

where

$$b_{1j} = \frac{a_{1j}}{a_{11}} \quad (j > 1)$$

Using equation (2), it is easy to eliminate the unknown x_1 from the system (1). To do this, subtract (2) multiplied by a_{21} from the second equation of (1), subtract (2) multiplied by a_{31} from the third equation of (1), and so on. We finally get a system consisting of three equations:

$$\left.\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 = a_{25}^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 = a_{35}^{(1)}, \\ a_{42}^{(1)}x_2 + a_{43}^{(1)}x_3 + a_{44}^{(1)}x_4 = a_{45}^{(1)} \end{aligned}\right\} \tag{1'}$$

where the coefficients $a_{ij}^{(1)}$ $(i, j \geqslant 2)$ are computed from the formula

$$a_{ij}^{(1)} = a_{ij} - a_{i1}b_{1j} \quad (i, j \geqslant 2).$$

Now, dividing the coefficients of the first equation of (1′) by the **leading element** $a_{22}^{(1)}$, we get the equation

$$x_2 + b_{23}^{(1)}x_3 + b_{24}^{(1)}x_4 = b_{25}^{(1)} \tag{2'}$$

where

$$b_{2j}^{(1)} = \frac{a_{2j}^{(1)}}{a_{22}^{(1)}} \quad (j > 2)$$

Eliminating x_2 in the same way that we eliminated x_1, we arrive at the following system of equations:

$$\left.\begin{aligned} a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 = a_{35}^{(2)}, \\ a_{43}^{(2)}x_3 + a_{44}^{(2)}x_4 = a_{45}^{(2)} \end{aligned}\right\} \tag{1''}$$

where

$$a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2}^{(1)}b_{2j}^{(1)} \quad (i, j \geqslant 3)$$

Dividing the coefficients of the first equation of (1″) by the **leading element** $a_{33}^{(2)}$, we get

$$x_3 + b_{34}^{(2)}x_4 = b_{35}^{(2)} \tag{2''}$$

where

$$b_{3j}^{(2)} = \frac{a_{3j}^{(2)}}{a_{33}^{(2)}} \quad (j > 3)$$

Now, eliminating x_3 in the same fashion from (1″), we get

$$a_{44}^{(3)}x_4 = a_{45}^{(3)} \tag{1'''}$$

where

$$a_{ij}^{(3)} = a_{ij}^{(2)} - a_{i3}^{(2)} b_{3j}^{(2)} \qquad (i,\ j \geqslant 4)$$

whence

$$x_4 = \frac{a_{45}^{(3)}}{a_{44}^{(3)}} = b_{45}^{(3)} \tag{2'''}$$

The remaining unknowns are successively determined from the equations (2″), (2′) and (2):

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4,$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3,$$

$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2$$

Thus, the process of solving a linear system of equations by the Gaussian method reduces to the construction of an equivalent system (2), (2′), (2″), (2‴) having a triangular matrix. A necessary and sufficient condition for using this method is that all the leading elements be nonzero. The computations can be conveniently arranged as shown in Table 13. The scheme in the table is called the *scheme of unique division*. The process of finding the coefficients $b_{ij}^{(i-1)}$ of a triangular system may be called *forward substitution (direct procedure)*, the process of finding the values of the unknowns is called *back substitution (reverse procedure)*.

The direct procedure begins with writing out the coefficients of the system, including the constant terms (Section A). The last row of Section A is the result of dividing the first row of the section by the leading element a_{11}. The elements $a_{ij}^{(1)}$ $(i,\ j \geqslant 2)$ of the next section of the scheme (Section A_1) are equal to the corresponding elements a_{ij} of the preceding section minus the product of their "projections" by the lines of Section A containing element 1 (that is to say, by the first column and last row).

The last row of Section A_1 is found by means of dividing the first row of the section by the leading element $a_{22}^{(1)}$. The subsequent sections are constructed in similar fashion. The direct procedure is terminated when we reach the section consisting of one row, not counting the transformed row (Section A_3 in our particular case).

In back substitution (the reverse procedure), use is made only of the rows of sections A_i containing units (*marked rows*), beginning with the last. The element $b_{45}^{(3)}$ of Section A_3 in the column of constant terms of the marked row of the section yields the value of x_4. From then on, all the other unknowns x_i $(i = 3,\ 2,\ 1)$ are found step by step by means of subtracting from the constant term

TABLE 13

SCHEME OF UNIQUE DIVISION

x_1	x_2	x_3	x_4	Constant terms	$\sum$	Sections of scheme
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	A
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	
1	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}	
	$a^{(1)}_{22}$	$a^{(1)}_{23}$	$a^{(1)}_{24}$	$a^{(1)}_{25}$	$a^{(1)}_{26}$	A_1
	$a^{(1)}_{32}$	$a^{(1)}_{33}$	$a^{(1)}_{34}$	$a^{(1)}_{35}$	$a^{(1)}_{36}$	
	$a^{(1)}_{42}$	$a^{(1)}_{43}$	$a^{(1)}_{44}$	$a^{(1)}_{45}$	$a^{(1)}_{46}$	
*	1	$b^{(1)}_{23}$	$b^{(1)}_{24}$	$b^{(1)}_{25}$	$b^{(1)}_{26}$	
		$a^{(2)}_{33}$	$a^{(2)}_{34}$	$a^{(2)}_{35}$	$a^{(2)}_{36}$	A_2
		$a^{(2)}_{43}$	$a^{(2)}_{44}$	$a^{(2)}_{45}$	$a^{(2)}_{46}$	
		1	$b^{(2)}_{34}$	$b^{(2)}_{35}$	$b^{(2)}_{36}$	
			$a^{(3)}_{44}$	$a^{(3)}_{45}$	$a^{(3)}_{46}$	A_3
			1	$b^{(3)}_{45}$	$b^{(3)}_{46}$	
				(x_4)	$(\bar{x}_4)$	
			1	x_4	$\bar{x}_4$	B
		1		x_3	$\bar{x}_3$	
	1			x_2	$\bar{x}_2$	
1				x_1	$\bar{x}_1$	

of the marked row the sums of the products of its coefficients by the corresponding values of the earlier found unknowns. The values of the unknowns are written out in succession in the last section, Section B. The units in this section help to locate for x_i the respective coefficients in the marked rows.

The computations are checked by so-called "check sums"

$$a_{i6}=\sum_{j=1}^{5} a_{ij} \qquad (i=1,\ 2,\ \ldots,\ 5) \tag{3}$$

which are located in the column labelled $\sum$ and are the sums of the elements of the rows of the matrix of the original system (1), including the constant terms.

If a_{i6} is taken as the new constant terms in system (1), then the transformed linear system

$$\sum_{j=1}^{4} a_{ij}\bar{x}_j=a_{i6} \qquad (i=1,\ 2,\ 3,\ 4) \tag{4}$$

will have the unknowns $\bar{x}_j$ connected with the earlier unknowns x_j by the relations

$$\bar{x}_j=x_j+1 \qquad (j=1,\ 2,\ 3,\ 4) \tag{5}$$

Indeed, substituting formulas (5) into equation (4), we get, by virtue of the system (1) and formulas (3), the identity

$$\sum_{j=1}^{4} a_{ij}x_j+\sum_{j=1}^{4} a_{ij}=\sum_{j=1}^{5} a_{ij}\equiv a_{i6} \qquad (j=1,\ 2,\ 3,\ 4)$$

Generally, if we perform the same operations with the check sums in each row as with the remaining elements of that row, then in the absence of errors in the computations, the elements of the column headed $\sum$ will be equal to the sums of the elements of the corresponding transformed rows. This serves as a check on the direct procedure. The reverse procedure is checked by finding the numbers $\bar{x}_j$, which must coincide with the numbers x_j+1.

Example. Solve the system

$$\left.\begin{aligned}
7.9x_1+5.6x_2+5.7x_3-7.2x_4&=6.68,\\
8.5x_1-4.8x_2+0.8x_3+3.5x_4&=9.95,\\
4.3x_1+4.2x_2-3.2x_3+9.3x_4&=8.6,\\
3.2x_1-1.4x_2-8.9x_3+3.3x_4&=1
\end{aligned}\right\} \tag{6}$$

Solution. In Section A of Table 14 write down the matrix of the coefficients of the system, its constant terms and the check sums. Then fill in the last (fifth) row of Section A, dividing the first row by 7.9 (by a_{11}).

TABLE 14

SOLUTION OF A SYSTEM OF EQUATIONS BY THE SCHEME OF UNIQUE DIVISION

x_1	x_2	x_3	x_4	Constant terms	$\sum$	Sections of scheme
7.9	5.6	5.7	—7.2	6.68	18.68	A
8.5	—4.8	0.8	3.5	9.95	17.95	
4.3	4.2	—3.2	9.3	8.6	23.2	
3.2	—1.4	—8.9	3.3	1	—2.8	
1	0.70886	0.72152	—0.91139	0.84557	2.36456	
	—10.82531	—5.33292	11.24682	2.76265	—2.14876	A_1
	1.15190	—6.30254	13.21898	4.96405	13.03239	
	—3.66835	—11.20886	6.21645	—1.70582	—10.36658	
	1	0.49263	—1.03894	—0.25520	0.19849	
		—6.87000	14.41573	5.25801	12.80374	A_2
		—9.40172	2.40525	—2.64198	—9.63845	
		1	—2.09836	—0.76536	—1.86372	
			—17.32294	—9.83768	—27.16062	A_3
			1	0.56790	1.56790	
			1	0.56790	1.56790	B
		1		0.42630	1.42630	
	1			0.12480	1.12480	
1				0.96710	1.96710	

Then start filling in Section A_1 of the table. Taking any element of Section A (not in the first row), subtract from it the product of the first element of its row by the last element of the column it belongs to, and record it in the appropriate place in Section A_1 of the scheme. For instance, choosing $a_{43} = -8.9$, we have

$$a_{43}^{(1)} = a_{43} - a_{41} b_{13} = -8.9 - 3.2 \cdot 0.72152 = -11.20886$$

To obtain the last row of Section A_1, divide all terms of the first row of that section by $a_{22}^{(1)} = -10.82531$. For example,

$$b_{23}^{(1)} = \frac{a_{23}^{(1)}}{a_{22}^{(1)}} = \frac{-5.33292}{-10.82531} = 0.49263$$

All the remaining sections of the table are filled in similarly. For instance,

$$a_{44}^{(2)} = a_{44}^{(1)} - a_{42}^{(1)} b_{24}^{(1)} = 6.21645 - (-3.66835) \cdot (-1.03894) = 2.40525$$

The unknowns are found by using the rows containing units, beginning with the last one (*marked rows*). The unknown x_4 is the constant term of the last row of Section A_3:

$$x_4 = b_{45}^{(3)} = 0.56790$$

The values of the other unknowns x_3, x_2, x_1 are obtained in succession by subtracting from the constant terms of the marked rows the sums of the products of the corresponding coefficients $b_{ij}^{(k)}$ by the earlier found values of the unknowns.

We have

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4 = -0.76536 - (-2.09836) \cdot 0.56790 = 0.42630,$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3 =$$
$$= -0.25520 - (-1.03894) \cdot 0.56790 - 0.49263 \cdot 0.42630 = 0.12480,$$

$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2 = 0.84557 - (-0.91139) \times$$
$$\times 0.56790 - 0.72152 \cdot 0.42630 - 0.70886 \cdot 0.12480 = 0.96710$$

thus,

$$x_1 = 0.96710, \quad x_2 = 0.12480, \quad x_3 = 0.42630, \quad x_4 = 0.56790$$

An intermediate check of the computations is carried out by means of the $\sum$ column on which are performed all the operations performed on the other columns.

Thus, (1) the sum of the elements of each row of the scheme (the elements not belonging to the $\sum$ column) must be equal to the element of that row of the $\sum$ column; (2) the roots $\bar{x}_i$ that correspond to the $\sum$ column must be greater by unity than the corresponding roots of the system.

Incidentally, if we take into account the units written in Section B, then again the elements of the $\sum$ column in this section are sums of the elements of the rows corresponding to them. In our case, the first and second conditions are valid to within unity of the last digit place. It is therefore almost definite that the computations were performed correctly.

Note that if the matrix of the system is symmetric, the corresponding parts of the sections A, A_1, A_2, ... of the scheme of unique division are also symmetric. This circumstance can be utilized to simplify the table.

It is easy to estimate the number of arithmetic operations, N, necessary to solve a system of linear equations in n unknowns by the Gaussian method [5] (not counting those for checking).

The direct procedure requires the following number of multiplications and divisions:

$$n(n+1)+(n-1)n+\ldots+1\cdot 2=$$

$$=(1^2+2^2+\ldots+n^2)+(1+2+\ldots+n)=\frac{n(n+1)(n+2)}{3}$$

and as many subtractions. The reverse procedure requires $\frac{n(n-1)}{2}$ multiplications and divisions and the same number of subtractions. Hence, the total number of arithmetic operations in the Gaussian method is

$$N=\frac{2n(n+1)(n+2)}{3}+n(n-1)<n^3$$

for $n>7$.

Thus, the time required for the solution of a linear system by the Gaussian method is roughly proportional to the cube of the number of unknowns. For example, to solve a system of 100 linear equations in 100 unknowns by the Gaussian method on a computer capable of 10^4 operations per second requires

$$T=10^6\cdot 10^{-4}=100 \text{ seconds}$$

The actual machine time will be considerably greater because of other operations in the routine besides arithmetic operations (address substitution, logical operations, sending, shaping, etc.).

8.4 IMPROVING ROOTS

Approximate values of roots obtained by the Gaussian method can be improved. We will show how this is done if the corrections to the roots are small in absolute value.

Suppose an approximate solution $\boldsymbol{x}_0$ is found for the system

$$A\boldsymbol{x}=\boldsymbol{b}$$

Setting

$$\boldsymbol{x}=\boldsymbol{x}_0+\boldsymbol{\delta}$$

we then get, for the correction $\boldsymbol{\delta}=\begin{bmatrix}\delta_1\\ \cdot\\ \cdot\\ \cdot\\ \delta_n\end{bmatrix}$ of the root $\boldsymbol{x}_0$,

$$A(\boldsymbol{x}_0+\boldsymbol{\delta})=\boldsymbol{b}$$

or

$$A\delta = \varepsilon$$

where $\varepsilon = \boldsymbol{b} - A\boldsymbol{x}_0$ is the *residual of the approximate solution* $\boldsymbol{x}_0$. Thus, in order to find δ, it is necessary to solve a linear system with the earlier matrix A and a new constant term $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$. To do this, all we need is to adjoin to the main computational scheme a column ε of constant terms and transform it by the general rules. As usual, the corrections $\delta_1, \delta_2, \ldots, \delta_n$ are found from the marked rows, the coefficients of these unknown corrections being already given in the table. Note that the transformed coefficients of matrix A need not be improved since for small residuals the corresponding errors will have a higher order of smallness.

Example. Solve the following system of equations by the Gaussian method to three places (say, by slide rule or hand):

$$\left.\begin{aligned} 6x_1 - x_2 - x_3 &= 11.33, \\ -x_1 + 6x_2 - x_3 &= 32, \\ -x_1 - x_2 + 6x_3 &= 42 \end{aligned}\right\} \qquad (1)$$

Using the values thus obtained as initial approximations, improve the roots to 10^{-4}.

Solution. Use the ordinary scheme of unique division (Table 15) and carry out all operations with three significant digits.

The approximate values of the roots are:

$$x_1^{(0)} = 4.67, \quad x_2^{(0)} = 7.62, \quad x_3^{(0)} = 9.05$$

Substituting these values in the given system (1), we compute the appropriate residuals [that is, the differences between the right and left members of system (1)]:

$$\varepsilon_1^{(0)} = -0.02, \quad \varepsilon_2^{(0)} = 0, \quad \varepsilon_3^{(0)} = -0.01$$

Using these values as constant terms (Table 15), we obtain the corrections of the roots:

$$\delta_1^{(0)} = -0.0039, \quad \delta_2^{(0)} = -0.0011, \quad \delta_3^{(0)} = -0.0025$$

whence we get the improved values of the roots:

$$x_1 = 4.6661, \quad x_2 = 7.6189, \quad x_3 = 9.0475$$

the residuals being equal to

$$\delta_1 = -2 \cdot 10^{-4}, \quad \delta_2 = 2 \cdot 10^{-4}, \quad \delta_3 = 0$$

TABLE 15

REFINING ROOTS COMPUTED BY THE GAUSSIAN METHOD

x_1	x_2	x_3	Constant terms	$\sum$	Residual ε
6	—1	—1	11.33	15.33	—0.02
—1	6	—1	32	36	0
—1	—1	6	42	46	—0.01
1	—0.167	—0.167	1.89	2.56	—0.0033
	5.83	—1.17	33.9	38.6	—0.0033
	—1.17	5.83	43.9	48.6	—0.0133
	1	—0.200	5.80	6.60	—0.0006
		5.60	50.7	56.3	—0.0140
		1	9.05 9.0475	10.05	—0.0025
	1		7.62 7.6189	8.62	—0.0011
1			4.67 4.6661		—0.0039

It is sometimes required to determine a possible error $\Delta \boldsymbol{x}$ of the root $\boldsymbol{x}$ of a linear system on the basis of known small errors ΔA and $\Delta \boldsymbol{b}$ of the matrix A of the system and its constant term $\boldsymbol{b}$.

We have

$$A\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

and

$$(A + \Delta A)(\boldsymbol{x} + \Delta \boldsymbol{x}) = \boldsymbol{b} + \Delta \boldsymbol{b} \tag{3}$$

From this, neglecting the small term $\Delta A \cdot \Delta \boldsymbol{x}$, we obtain

$$A\boldsymbol{x} + A\Delta \boldsymbol{x} + \Delta A \boldsymbol{x} = \boldsymbol{b} + \Delta \boldsymbol{b}$$

or

$$A\Delta \boldsymbol{x} = \Delta \boldsymbol{b} - \Delta A \boldsymbol{x} \tag{4}$$

It is thus possible, when seeking $\Delta\boldsymbol{x}$ approximately, to use the Gaussian scheme for the basic system (2) by augmenting the scheme with a new column of constant terms, $\Delta\boldsymbol{b}-\Delta A\boldsymbol{x}$.

8.5 THE METHOD OF PRINCIPAL ELEMENTS

Suppose we have a linear system

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=a_{1,\,n+1}, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=a_{2,\,n+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=a_{n,\,n+1} \end{array}\right\} \qquad (1)$$

Consider the augmented rectangular matrix consisting of the coefficients of the system and its constant terms,

$$M=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1q} & \ldots & a_{1n} & a_{1,\,n+1} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2q} & \ldots & a_{2n} & a_{2,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & a_{i2} & \ldots & a_{ij} & \ldots & a_{iq} & \ldots & a_{in} & a_{i,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & a_{p2} & \ldots & a_{pj} & \ldots & \boxed{a_{pq}} & \ldots & a_{pn} & a_{p,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nj} & \ldots & a_{nq} & \ldots & a_{nn} & a_{n,\,n+1} \end{bmatrix}$$

Choose a nonzero (as a rule, the numerically largest) element a_{pq} of matrix M not belonging to the column of constant terms $(q\neq n+1)$, this element being called the *principal element*. Compute the multipliers

$$m_i=-\frac{a_{iq}}{a_{pq}}$$

for all $i\neq p$.

The row of M with index p which contains the principal element is called the *principal row*. Then perform the following operation: to each nonprincipal row add the principal row multiplied by the appropriate multiplier m_i of the row. We thus obtain a new matrix in which the qth column consists of zeros. Discarding this column and the principal pth row, we obtain a new matrix $M^{(1)}$ with the number of rows and columns diminished by unity.

Repeat these operations with matrix $M^{(1)}$ to get matrix $M^{(2)}$, etc. Thus, we obtain a sequence of matrices

$$M,\ M^{(1)},\ \ldots,\ M^{(n-1)}$$

the last of which is a two-term row matrix. It is also regarded as the principal row.

To determine the unknowns x_i, combine into a system all the principal rows, beginning with the last, which enters into matrix $M^{(n-1)}$.

After an appropriate change in the numbering of the unknowns we get a system with a triangular matrix, from which it is easy to obtain, step by step, the unknowns of the given system (1). The method of principal elements is always applicable if the determinant of the system

$$\det A = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \neq 0$$

Note that the Gaussian method is a particular case of the method of principal elements and the Gaussian scheme is obtained if for the principal element we always choose the upper left element of the corresponding matrix.

8.6 USE OF THE GAUSSIAN METHOD IN COMPUTING DETERMINANTS

Suppose

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot\ \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{1}$$

and

$$\Delta = \det A \tag{2}$$

Consider the linear system

$$A\boldsymbol{x} = 0 \tag{3}$$

When solving (3) by the Gaussian method we replaced matrix A by the triangular matrix B consisting of elements of marked rows,

$$B = \begin{bmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & 1 & b_{23}^{(1)} & \dots & b_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot\ \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

We obtained an equivalent system:

$$B\boldsymbol{x} = 0 \tag{4}$$

The elements of B were successively obtained from the elements of A and the subsequent auxiliary matrices $A_1, A_2, \ldots, A_{n-1}$ with the aid of the following elementary transformations:

(1) division by the leading elements $a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$, which were assumed nonzero, and

(2) subtraction, from the rows of A and the intermediate matrices A_i $(i = 1, 2, \ldots, n-1)$, of scalars proportional to the elements of the corresponding leading rows. In the first operation, the determinant of the matrix is also divided by the appropriate leading element, in the second, the determinant of the matrix remains unchanged. Therefore

$$\det B = 1 = \frac{\det A}{a_{11}a_{22}^{(1)}\ldots a_{nn}^{(n-1)}}$$

Hence

$$\Delta = \det A = a_{11}a_{22}^{(1)}\ldots a_{nn}^{(n-1)} \tag{5}$$

that is, *the determinant is equal to the product of the leading elements of the corresponding Gaussian scheme.* We conclude from this that the scheme of unique division given in Sec. 8.3 may be used for computing determinants (in this case the column of constant terms is superfluous).

Note that if for any step the element $a_{ii}^{(i-1)} = 0$ or is close to zero (this implies a reduction in the accuracy of the computations), one should appropriately change the order of the rows and columns of the matrix.

Example. Evaluate the determinant

$$\Delta = \begin{vmatrix} 7.4 & 2.2 & -3.1 & 0.7 \\ 1.6 & 4.8 & -8.5 & 4.5 \\ 4.7 & 7.0 & -6.0 & 6.6 \\ 5.9 & 2.7 & 4.9 & -5.3 \end{vmatrix}$$

Solution. Using the elements of the determinant Δ, form a unique-division scheme (Table 16).

Multiplying together the leading elements (in frames), we get

$$\Delta = 7.4 \cdot 4.32434 \cdot 6.11331 \cdot (-7.58393) = -1483.61867$$

It is worth noting the following. To solve a system of n linear equations in n unknowns by Cramer's formulas, one has to evaluate $n+1$ determinants of the nth order. Now by the unique-division scheme, to compute one determinant of the nth order requires nearly the same volume of work as the complete solution of the system of equations. It is therefore, generally speaking, not advisable to use Cramer's rule for a numerical solution of a linear system of equations for $n > 3$.

TABLE 16

EVALUATING A DETERMINANT BY THE GAUSSIAN METHOD

1st column	2nd column	3rd column	4th column	Σ	
$\boxed{7.4}$	2.2	−3.1	0.7	7.2	A
1.6	4.8	−8.5	4.5	2.4	
4.7	7.0	−6.0	6.6	12.3	
5.9	2.7	4.9	−5.3	8.2	
1	0.29729	−0.41891	0.09459	0.97297	
	$\boxed{4.32434}$	−7.82974	4.34866	0.84326	A_1
	5.60274	−4.03112	6.15543	7.72705	
	0.94599	7.37157	−5.85808	2.45948	
	1	−1.81062	1.00562	0.19500	
		$\boxed{6.11331}$	0.52120	6.63451	A_2
		9.08440	−6.80939	2.27501	
		1	0.08526	1.08526	
			$\boxed{-7.58393}$	−7.58393	A_3
				$\Delta = -1483.61867$	

8.7 INVERSION OF MATRICES BY THE GAUSSIAN METHOD

Suppose we have a nonsingular matrix

$$A = [a_{ij}] \qquad (i,\ j = 1,\ 2,\ \ldots,\ n) \tag{1}$$

To find its inverse

$$A^{-1} = [x_{ij}] \tag{2}$$

we use the basic relation

$$AA^{-1}=E \tag{3}$$

where E is the unit matrix.

Multiplying matrices A and A^{-1}, we get n systems of equations in n^2 unknowns x_{ij}:

$$\sum_{k=1}^{n} a_{ik}x_{kj}=\delta_{ij} \qquad (i,\ j=1,\ 2,\ \ldots,\ n)$$

where

$$\delta_{ij}=\begin{cases} 1 & \text{when} \quad i=j, \\ 0 & \text{when} \quad i\neq j \end{cases}$$

The resulting n systems of linear equations for $j=1, 2, \ldots, n$ having the same matrix A and distinct constant terms can be solved simultaneously by the Gaussian method.

Example. Find the inverse A^{-1} of the matrix

$$A=\begin{bmatrix} 1.8 & -3.8 & 0.7 & -3.7 \\ 0.7 & 2.1 & -2.6 & -2.8 \\ 7.3 & 8.1 & 1.7 & -4.9 \\ 1.9 & -4.3 & -4.9 & -4.7 \end{bmatrix}$$

Solution. Form a unique-division scheme. We will have four columns of constant terms (Table 17). Note that the elements of the rows of the inverse matrix are obtained in reverse order.

On the basis of the results of Table 17, we get

$$A^{-1}=\begin{bmatrix} -0.21121 & -0.46003 & 0.16284 & 0.26956 \\ -0.03533 & 0.16873 & 0.01573 & -0.08920 \\ 0.23030 & 0.04607 & -0.00944 & -0.19885 \\ -0.29316 & -0.38837 & 0.06128 & 0.18513 \end{bmatrix}$$

To check, form the product

$$AA^{-1}=\begin{bmatrix} 1.8 & -3.8 & 0.7 & -3.7 \\ 0.7 & 2.1 & -2.6 & -2.8 \\ 7.3 & 8.1 & 1.7 & -4.9 \\ 1.9 & -4.3 & -4.9 & -4.7 \end{bmatrix}\times$$

$$\times\begin{bmatrix} -0.21121 & -0.46003 & 0.16284 & 0.26956 \\ -0.03533 & 0.16873 & 0.01573 & -0.08920 \\ 0.23030 & 0.04607 & -0.00944 & -0.19885 \\ -0.29316 & -0.38837 & 0.06128 & 0.18513 \end{bmatrix}=$$

TABLE 17

COMPUTING THE INVERSE MATRIX BY THE GAUSSIAN METHOD

x_{1j}	x_{2j}	x_{3j}	x_{4j}	$j=1$	$j=2$	$j=3$	$j=4$	Σ
1.8	—3.8	0.7	—3.7	1	0	0	0	—4.0
0.7	2.1	—2.6	—2.8	0	1	0	0	—1.6
7.3	8.1	1.7	—4.9	0	0	1	0	13.2
1.9	—4.3	—4.9	—4.7	0	0	0	1	—11.0
1	—2.11111	0.38889	—2.05556	0.55556	0	0	0	—2.22223
	3.57778	—2.87222	—1.36111	—0.38885	1	0	0	—0.04440
	23.51110	—1.13890	10.10559	—4.05551	0	1	0	29.42228
	—0.28889	—5.63889	—0.79444	—1.05554	0	0	1	—6.77776
	1	—0.80279	—0.38043	—0.10868	0.27950	0	0	—0.01241
		17.73557	19.04992	—1.50032	—6.57135	1	0	29.71405
		—5.87081	—0.90434	—1.08694	0.08074	0	1	—6.78134
		1	1.07411	—0.08459	—0.37108	0.05638	0	1.67539
			5.40155	—1.58355	—2.09780	0.33100	1	3.05456
			1	—0.29316	—0.38837	0.06128	0.18513	0.56540
		1		0.23030	0.04607	—0.00944	—0.19885	1.06809
	1			—0.03533	0.16873	0.01573	0.08920	1.06013
1				—0.21121	—0.46003	0.16284	0.26956	0.76266

$$=\begin{bmatrix} 0.99997 & 0.00000 & -0.00001 & 0.00000 \\ -0.00025 & 0.99997 & -0.00002 & -0.00039 \\ -0.00808 & -0.01017 & 0.99982 & 0.00009 \\ 0.00000 & 0.00000 & 0.00000 & 1.00048 \end{bmatrix}=$$

$$=E-10^{-3}\cdot\begin{bmatrix} 0.03 & 0.00 & 0.01 & 0.00 \\ 0.25 & 0.03 & 0.02 & 0.39 \\ 8.08 & 10.17 & 0.18 & -0.09 \\ 0.00 & 0.00 & 0.00 & -0.48 \end{bmatrix}$$

We see that due to rounding the inverse is not quite exact. Below we give (see Sec. 8.15) a method for correcting the elements of an approximate inverse matrix.

8.8 SQUARE-ROOT METHOD

Suppose we have a linear system

$$A\boldsymbol{x}=\boldsymbol{b} \tag{1}$$

where $A=[a_{ij}]$ is a symmetric matrix, that is $A'=[a_{ji}]=A$. Then A may be given in the form of a product of two transposed triangular matrices:

$$A=T'T \tag{2}$$

where

$$T=\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & t_{nn} \end{bmatrix} \quad \text{and} \quad T'=\begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{12} & t_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ t_{1n} & t_{2n} & \dots & t_{nn} \end{bmatrix}$$

Multiplying together the matrices T' and T, we obtain the following equations which allow us to determine the elements t_{ij} of matrix T:

$$\left.\begin{aligned} t_{1i}t_{1j}+t_{2i}t_{2j}+\dots+t_{ii}t_{ij}&=a_{ij} \qquad (i<j), \\ t_{1i}^2+t_{2i}^2+\dots+t_{ii}^2&=a_{ii} \end{aligned}\right\}$$

Whence we find in succession:

$$\left.\begin{aligned} & t_{11}=\sqrt{a_{11}},\ t_{1j}=\frac{a_{1j}}{t_{11}} \qquad (j>1), \\ & t_{ii}=\sqrt{a_{ii}-\sum_{k=1}^{i-1}t_{ki}^2} \qquad (1<i\leqslant n), \\ & t_{ij}=\frac{a_{ij}-\sum_{k=1}^{i-1}t_{ki}t_{kj}}{t_{ii}} \qquad (i<j), \\ & t_{ij}=0 \quad \text{for} \quad i>j \end{aligned}\right\} \tag{3}$$

The system (1) has a definite unique solution if $t_{ii} \neq 0$ ($i = 1, 2, \ldots, n$), since in that case

$$\det A = \det T' \cdot \det T = (\det T)^2 = (t_{11} t_{22} \ldots t_{nn})^2 \neq 0$$

The coefficients of matrix T will be real if $t_{ii}^2 > 0$. In the sequel we will not, generally speaking, assume this latter condition to be fulfilled.

Given relation (2), the equation (1) is equivalent to two equations:

$$T'\boldsymbol{y} = \boldsymbol{b} \quad \text{and} \quad T\boldsymbol{x} = \boldsymbol{y}$$

or, expanded,

$$\left.\begin{aligned} &t_{11}y_1 = b_1, \\ &t_{12}y_1 + t_{22}y_2 = b_2, \\ &\cdots\cdots\cdots\cdots \\ &t_{1n}y_1 + t_{2n}y_2 + \ldots + t_{nn}y_n = b_n \end{aligned}\right\} \tag{4}$$

and

$$\left.\begin{aligned} t_{11}x_1 + t_{12}x_2 + \ldots + t_{1n}x_n &= y_1, \\ t_{22}x_2 + \ldots + t_{2n}x_n &= y_2, \\ \cdots\cdots\cdots\cdots\cdots & \\ t_{nn}x_n &= y_n \end{aligned}\right\} \tag{5}$$

From this we successively obtain

$$\left.\begin{aligned} &y_1 = \frac{b_1}{t_{11}}, \\ &y_i = \frac{b_i - \sum_{k=1}^{i-1} t_{ki}y_k}{t_{ii}} \quad (i > 1) \end{aligned}\right\} \tag{6}$$

and

$$\left.\begin{aligned} &x_n = \frac{y_n}{t_{nn}}, \\ &x_i = \frac{y_i - \sum_{k=i+1}^{n} t_{ik}x_k}{t_{ii}} \quad (i < n) \end{aligned}\right\} \tag{7}$$

The foregoing method of solving a system of linear equations is called the *square-root method*. Since matrix A is symmetric and matrix T is upper triangular, we need only write $\frac{n}{2}(n+1)$ upper coefficients a_{ij} and $t_{ij} (i \geqslant j)$ in the computational scheme. The checking procedure is the ordinary type with the aid of sums; all coefficients of the appropriate row are taken into account when forming a sum.

Note that if for some sth row we have $t_{ss}^2 < 0$, then the corresponding elements t_{sj} will be imaginary. Formally, the method is applicable in this case as well.

In practical applications of the square-root method, the *direct procedure*, by means of formulas (3) and (6), is used to compute successively the coefficients t_{ij} and $y_i (i = 1, 2, \ldots, n)$ and then the *reverse procedure*, by formula (7), is used to find the unknowns $x_i (i = n, n-1, \ldots, 1)$.

Example. Using the square-root method, solve the following system of equations:

$$\left.\begin{aligned} x_1 + 3x_2 - 2x_3 \qquad - 2x_5 &= 0.5, \\ 3x_1 + 4x_2 - 5x_3 + x_4 - 3x_5 &= 5.4, \\ -2x_1 - 5x_2 + 3x_3 - 2x_4 + 2x_5 &= 5.0, \\ x_2 - 2x_3 + 5x_4 + 3x_5 &= 7.5, \\ -2x_1 - 3x_2 + 2x_3 + 3x_4 + 4x_5 &= 3.3 \end{aligned}\right\}$$

Solution. Enter the coefficients a_{ij} and the constant terms b_i of the given system in the initial section, A, of the table (Table 18)

TABLE 18

SOLUTION OF A LINEAR SYSTEM BY THE SQUARE-ROOT METHOD

a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	b_i	Σ	Sections of scheme
1	3	−2	0	−2	0.5	0.5	*A*
3	4	−5	1	−3	5.4	5.4	
−2	−5	3	−2	2	5.0	1.0	
0	1	−2	5	3	7.5	14.5	
−2	−3	2	3	4	3.3	7.3	
t_{i1}	t_{i2}	t_{i3}	t_{i4}	t_{i5}	y_i	Σ	
1	3	−2	0	−2	0.5	0.5	*B*
	2.2361i	−0.4472i	−0.4472i	−1.3416i	−1.7471i	−1.7471i	
		0.8944i	2.0125i	1.5653i	−7.5803i	−3.1081i	
			3.0414	2.2194	−2.2928	2.9679	
				0.8221i	0.1643i	0.9859i	
−6.0978	−2.2016	−6.8011	−0.8996	0.1998		x_i	*C*
−5.0973	−1.2017	−5.8004	0.1007	1.1992		$\bar{x}_i$	

and compute the column marked $\sum$. Using formulas (3) and (6) and moving from row to row in succession, compute the coefficients t_{ij} and the new constant terms y_i, thus filling Section B of the table.

For example,

$$t_{35}=\frac{a_{35}-t_{13}t_{15}-t_{23}t_{25}}{t_{33}}=\frac{2-(-2)(-2)-(-0.4472i)(-1.3416i)}{0.8944i}=1.5653i$$

Compute the column labelled $\sum$ for a check. On the basis of formulas (7) we find the values of the unknowns x_i and the checking values $\overline{x}_i=x_i+1$, entering them in Section C. For example,

$$x_3=\frac{y_3-t_{35}x_5-t_{34}x_4}{t_{33}}=\frac{-7.5803i-1.5652i\cdot 0.1998-2.0125i\cdot(-0.8996)}{0.8944i}=$$

$$=-6.8011$$

8.9 THE SCHEME OF KHALETSKY

For the sake of convenience, we write the system of linear equations in matrix notation as

$$\boldsymbol{Ax}=\boldsymbol{b} \tag{1}$$

where $A=[a_{ij}]$ is a square matrix of order n and

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \boldsymbol{b}=\begin{bmatrix} a_{1,\,n+1} \\ \cdot\ \cdot\ \cdot \\ a_{n,\,n+1} \end{bmatrix}$$

are column vectors. Represent matrix A in the form of a product of a lower triangular matrix $B=[b_{ij}]$ and an upper triangular matrix $C=[c_{ij}]$ with unit diagonal; thus,

$$A=BC \tag{2}$$

where

$$B=\begin{bmatrix} b_{11} & 0 & \ldots & 0 \\ b_{21} & b_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot\ \cdot & \cdot \\ b_{n1} & b_{n2} & \ldots & b_{nn} \end{bmatrix} \quad \text{and} \quad C=\begin{bmatrix} 1 & c_{12} & \ldots & c_{1n} \\ 0 & 1 & \ldots & c_{2n} \\ \cdot & \cdot & \cdot\ \cdot & \cdot \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Then the elements b_{ij} and c_{ij} are determined from the formulas

$$\left.\begin{aligned} &b_{i1}=a_{i1}, \\ &b_{ij}=a_{ij}-\sum_{k=1}^{j-1} b_{ik}c_{kj} \qquad (i\geqslant j>1) \end{aligned}\right\} \tag{3}$$

and

$$\left.\begin{aligned} c_{1j} &= \frac{a_{1j}}{b_{11}}, \\ c_{ij} &= \frac{1}{b_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} b_{ik}c_{kj}\right) \qquad (1 < i < j) \end{aligned}\right\} \tag{4}$$

Whence the desired vector $\boldsymbol{x}$ may be computed from the chain of equations

$$B\boldsymbol{y} = \boldsymbol{b}, \quad C\boldsymbol{x} = \boldsymbol{y} \tag{5}$$

Systems (5) are readily solved since the matrices B and C are triangular:

$$\left.\begin{aligned} y_1 &= \frac{a_{1,\,n+1}}{b_{11}}, \\ y_i &= \frac{1}{b_{ii}}\left(a_{i,\,n+1} - \sum_{k=1}^{i-1} b_{ik}y_k\right) \qquad (i > 1) \end{aligned}\right\} \tag{6}$$

and

$$\left.\begin{aligned} x_n &= y_n, \\ x_i &= y_i - \sum_{k=i+1}^{n} c_{ik}x_k \quad (i < n) \end{aligned}\right\} \tag{7}$$

From formulas (6) it is evident that the numbers y_i are advantageously computed together with the coefficients c_{ij}. This method is known as the *scheme of Khaletsky*. This scheme uses the ordinary type of checking by means of sums.

Note that if the matrix A is symmetric, that is $a_{ij} = a_{ji}$, then

$$c_{ij} = \frac{b_{ji}}{b_{ii}} \quad (i < j)$$

Khaletsky's scheme is convenient for machine computation since in this case the operations of "accumulation" (3) and (4) may be carried out without recording the intermediate results.

Example. Solve the system

$$\left.\begin{aligned} 3x_1 + x_2 - x_3 + 2x_4 &= 6, \\ -5x_1 + x_2 + 3x_3 - 4x_4 &= -12, \\ 2x_1 \qquad + x_3 - x_4 &= 1, \\ x_1 - 5x_2 + 3x_3 - 3x_4 &= 3 \end{aligned}\right\}$$

Solution (see Table 19).

TABLE 19

	x_1	x_2	x_3	x_4		$\sum$	x_1	x_2	x_3	x_4		$\sum$
I	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	3	1	—1	2	6	11
	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	—5	1	3	—4	—12	—17
	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	2	0	1	—1	1	3
	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	1	—5	3	—3	3	—1
II	b_{11} 1	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	3 1	0.333333	—0.333333	0.666667	2	3.666667
	b_{21}	b_{22} 1	c_{23}	c_{24}	c_{25}	c_{26}	—5	2.666667 1	0.5	—0.25	—0.75	0.5
	b_{31}	b_{32}	b_{33} 1	c_{34}	c_{35}	c_{36}	2	—0.666667	2 1	—1.25	—1.75	—2
	b_{41}	b_{42}	b_{43}	b_{44} 1	c_{45}	c_{46}	1	—5.333333	6	2.5 1	3	4
III					y_1	x_1					2	1
					y_2	x_2					—0.75	—1
					y_3	x_3					—1.75	2
					y_4	x_4					3	3

In the first section of Table 19 enter the matrix of coefficients of the system, the constant terms and the check sums.

Then, since $b_{i1} = a_{i1}$ $(i = 1, 2, 3, 4)$, the first column of Section I is moved to the first column of Section II.

To obtain the first row of Section II, divide all elements of the first row of Section I by the element $a_{11} = b_{11}$, by 3 in our case. We have

$$c_{12} = \frac{1}{3} = 0.(3),$$

$$c_{13} = -\frac{1}{3} = -0.(3),$$

$$c_{14} = \frac{2}{3} = 0.(6),$$

$$c_{15} = \frac{6}{3} = 2,$$

$$c_{16} = \frac{11}{3} = 2.(6)$$

Now fill in the second column of Section II beginning with the second row. Using formulas (3), determine b_{j2}:

$$b_{22} = a_{22} - b_{21}c_{12} = 1 - \left(-5 \cdot \frac{1}{3}\right) = \frac{8}{3} = 2.66(6),$$

$$b_{32} = a_{32} - b_{31}c_{12} = 0 - 2 \cdot \frac{1}{3} = -\frac{2}{3} = 0.(6),$$

$$b_{42} = a_{42} - b_{41}c_{12} = -5 - 1 \cdot \frac{1}{3} = -5 \cdot \frac{1}{3} = -5.(3)$$

Then, determining c_{2j} $(j = 3, 4, 5, 6)$ by formulas (4), fill in the second row of Section II:

$$c_{23} = \frac{1}{b_{22}}(a_{23} - b_{21}c_{13}) = \frac{3}{8}\left[3 - (-5) \cdot \left(-\frac{1}{3}\right)\right] = \frac{1}{2},$$

$$c_{24} = \frac{1}{b_{22}}(a_{24} - b_{21}c_{14}) = \frac{3}{8}\left[(-4) - (-5) \cdot \frac{2}{3}\right] = -\frac{1}{4},$$

$$c_{25} = \frac{1}{b_{22}}(a_{25} - b_{21}c_{15}) = \frac{3}{8}\left[(-12) - (-5) \cdot 2\right] = -\frac{3}{4},$$

$$c_{26} = \frac{1}{b_{22}}(a_{26} - b_{21}c_{16}) = \frac{3}{8}\left[(-17) - (-5) \cdot \frac{11}{3}\right] = \frac{1}{2}$$

We now go to the third column and compute its elements b_{33} and b_{34} from formulas (3), and so on until we have filled in the whole of Section II. We thus get a staircase arrangement in Section II: column-row, column-row, etc.

In Section III, we determine y_i and x_i $(i = 1, 2, 3, 4)$ using formulas (6) and (7).

Intermediate checking is done by means of the $\sum$ column, which is involved in the same operations as is the column of constant terms.

8.10 THE METHOD OF ITERATION

When a linear system has a large number of unknowns, the Gaussian scheme, which yields an exact solution, becomes very unwieldy. Under such conditions it is sometimes more convenient to use approximate numerical methods for finding the roots of the system. One of these methods is the *method of iteration*, which we present here.

Given a linear system

$$\left.\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2,\\ \ldots\ldots\ldots\ldots\ldots&\\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n&=b_n \end{aligned}\right\} \tag{1}$$

Introducing the matrices

$$A=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n}\\ a_{21} & a_{22} & \ldots & a_{2n}\\ \ldots & \ldots & \ldots & \ldots\\ a_{n1} & a_{n2} & \ldots & a_{nn}\end{bmatrix},\qquad x=\begin{bmatrix} x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix},\qquad b=\begin{bmatrix} b_1\\ b_2\\ \vdots\\ b_n\end{bmatrix}$$

we can write system (1) briefly as a matrix equation:

$$Ax=b \tag{1'}$$

Assuming that the diagonal coefficients

$$a_{ii}\neq 0 \qquad (i=1,\ 2,\ \ldots,\ n)$$

we solve the first equation of (1) for x_1, the second for x_2, and so on. We then get the equivalent system

$$\left.\begin{aligned} x_1&=\beta_1+\alpha_{12}x_2+\alpha_{13}x_3+\ldots+\alpha_{1n}x_n,\\ x_2&=\beta_2+\alpha_{21}x_1+\alpha_{23}x_3+\ldots+\alpha_{2n}x_n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ x_n&=\beta_n+\alpha_{n1}x_1+\alpha_{n2}x_2+\ldots+\alpha_{n,\,n-1}x_{n-1} \end{aligned}\right\} \tag{2}$$

where

$$\beta_i=\frac{b_i}{a_{ii}},\qquad \alpha_{ij}=-\frac{a_{ij}}{a_{ii}}\quad \text{for}\quad i\neq j$$

and $\alpha_{ij}=0$ for $i=j$ $(i,\ j=1,\ 2,\ \ldots,\ n)$.

Introducing the matrices

$$\alpha=\begin{bmatrix}\alpha_{11} & \alpha_{12} & \dots & \alpha_{1n}\\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn}\end{bmatrix} \quad \text{and} \quad \beta=\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_n\end{bmatrix}$$

we can write (2) in matrix form:

$$\boldsymbol{x}=\beta+\alpha\boldsymbol{x} \qquad (2')$$

We will solve system (2) by the **method of successive approximations.** For the zeroth approximation we take, say, the column of constant $\boldsymbol{x}^{(0)}=\beta$.

Then we consecutively construct the column matrices

$$\boldsymbol{x}^{(1)}=\beta+\alpha\boldsymbol{x}^{(0)}$$

(first approximation)

$$\boldsymbol{x}^{(2)}=\beta+\alpha\boldsymbol{x}^{(1)}$$

(second approximation), and so forth.

Generally speaking, any $(k+1)$th approximation is computed from the formula

$$\boldsymbol{x}^{(k+1)}=\beta+\alpha\boldsymbol{x}^{(k)} \qquad (k=0,\ 1,\ 2,\ \dots) \qquad (3)$$

If the sequence of approximations $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k)}, \dots$ has a limit

$$\boldsymbol{x}=\lim_{k\to\infty}\boldsymbol{x}^{(k)}$$

then this limit is the solution of system (2). Indeed, passing to the limit in (3), we have

$$\lim_{k\to\infty}\boldsymbol{x}^{(k+1)}=\beta+\alpha\lim_{k\to\infty}\boldsymbol{x}^{(k)}$$

or

$$\boldsymbol{x}=\beta+\alpha\boldsymbol{x}$$

which is to say that the limiting vector $\boldsymbol{x}$ is the solution of system (2′) and, consequently, of system (1).

Let us write out the formulas of the approximations in full:

$$\left.\begin{array}{l} x_i^{(0)}=\beta_i,\\ x_i^{(k+1)}=\beta_i+\sum\limits_{j=1}^{n}\alpha_{ij}x_j^{(k)}\\ (\alpha_{ii}=0;\quad i=1,\ \dots,\ n;\quad k=0,\ 1,\ 2,\ \dots)\end{array}\right\} \qquad (3')$$

It will be noted that it is sometimes more advantageous to reduce system (1) to (2) so that the coefficients α_{ii} are not equal to zero. For instance, take the equation

$$1.02x_1-0.15x_2=2.7$$

In order to apply the method of successive approximations it is natural to write it in the form

$$x_1=2.7-0.02x_1+0.15x_2$$

Generally, having a system

$$\sum_{j=1}^{n} a_{ij}x_j=b_i \quad (i=1,\ 2,\ \ldots,\ n)$$

we can put

$$a_{ii}=a_{ii}^{(1)}+a_{ii}^{(2)}$$

where $a_{ii}^{(1)}\neq 0$. Then the given system is equivalent to the reduced system

$$x_i=\beta_i+\sum_{j=1}^{n}\alpha_{ij}x_j \quad (i=1,\ 2,\ \ldots,\ n)$$

where

$$\beta_i=\frac{b_i}{a_{ii}^{(1)}},\quad \alpha_{ii}=-\frac{a_{ii}^{(2)}}{a_{ii}^{(1)}},\quad \alpha_{ij}=-\frac{a_{ij}}{a_{ii}^{(1)}}\quad \text{for } i\neq j$$

For this reason, from now on we will not, generally speaking, assume that $\alpha_{ii}=0$.

The method of successive approximations given by the formula (3) or (3′) is called the *method of iteration*. The process of iteration (3) converges rapidly, that is, the number of approximations necessary to obtain the roots of (1) to a given accuracy is small if the elements of the matrix α are small in absolute value. In other words, successful use of the iteration process requires that the moduli of the diagonal coefficients of system (1) be large in comparison with the moduli of the nondiagonal coefficients of this system (here the constant terms are immaterial).

Example 1. Solve the system

$$\left.\begin{aligned} 4x_1+0.24x_2-0.08x_3&=8,\\ 0.09x_1+3x_2-0.15x_3&=9,\\ 0.04x_1-0.08x_2+4x_3&=20 \end{aligned}\right\}\qquad (4)$$

by the method of iteration.

Solution. Here, the diagonal coefficients 4, 3, 4 of the system considerably exceed the remaining coefficients of the unknowns.

Reduce this system to the normal form (2),

$$\left.\begin{aligned} x_1 &= 2-0.06x_2+0.02x_3, \\ x_2 &= 3-0.03x_1+0.05x_3, \\ x_3 &= 5-0.01x_1+0.02x_2 \end{aligned}\right\} \tag{5}$$

System (5) can be written in matrix form as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} + \begin{bmatrix} 0 & -0.06 & 0.02 \\ -0.03 & 0 & 0.05 \\ -0.01 & 0.02 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

For the zeroth approximations of the roots of (4) we take

$$x_1^{(0)}=2, \quad x_2^{(0)}=3, \quad x_3^{(0)}=5$$

Substituting these values into the right members of (5), we get the first approximations of the roots:

$$\begin{aligned} x_1^{(1)} &= 2-0.06\cdot 3+0.02\cdot 5=1.92, \\ x_2^{(1)} &= 3-0.03\cdot 2+0.05\cdot 5=3.19, \\ x_3^{(1)} &= 5-0.01\cdot 2+0.02\cdot 3=5.04 \end{aligned}$$

Substituting these approximations into (5), we get the second approximations of the roots:

$$x_1^{(2)}=1.9094, \quad x_2^{(2)}=3.1944, \quad x_3^{(2)}=5.0446$$

Substituting again, we get the third approximations of the roots:

$$x_1^{(3)}=1.90923, \quad x_2^{(3)}=3.19495, \quad x_3^{(3)}=5.04485, \text{ etc.}$$

The results of the computations are entered in Table 20.

TABLE 20

SOLVING A LINEAR SYSTEM BY THE METHOD OF ITERATION

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	2	2	5
1	1.92	3.19	5.04
2	1.9094	3.1944	5.0446
3	1.90923	3.19495	5.04485

Note. When using the method of iteration [formula (3)], it is not necessary to take the column of constant terms as the zeroth approximation $x^{(0)}$. As will be shown below, the convergence of the iteration process depends solely on the properties of the matrix α; note that when certain conditions are met, if this process converges for a certain choice of the initial approximation, it will converge to the same limiting vector for any other choice of the initial approximation as well. For this reason, in the iteration process the initial vector $x^{(0)}$ can be chosen **arbitrarily**. It is advisable, for the components of the initial vector, to take the approximate values of roots of the system found as a reasonable guess.

A converging iteration process is **self-correcting**; that is, an individual computational error will not affect the final result, since any erroneous approximation may be regarded as a new initial vector.

Also, it is sometimes more convenient to compute not the approximations as such but their differences. Introducing the notations

$$\Delta^{(k)} = x^{(k)} - x^{(k-1)} \qquad (k = 0, 1, 2, \ldots)$$

we have, from formula (3),

$$x^{(k+1)} = \beta + \alpha x^{(k)} \tag{6}$$

and

$$x^{(k)} = \beta + \alpha x^{(k-1)} \tag{7}$$

Whence subtract (7) from (6) to get

$$\Delta^{(k+1)} = \alpha\left(x^{(k)} - x^{(k-1)}\right) = \alpha\Delta^{(k)}$$

or

$$\Delta^{(k+1)} = \alpha\Delta^{(k)} \quad (k = 1, 2, \ldots) \tag{8}$$

For the zeroth approximation we take

$$\Delta^{(0)} = x^{(0)} \tag{9}$$

Then the mth approximation is

$$x^{(m)} = \sum_{k=0}^{m} \Delta^{(k)} \tag{10}$$

If, as usual, we put $\Delta^{(0)} = x^{(0)} = \beta$, then (8) will hold true for $k = 0$ as well, otherwise (8) does not hold for $k = 0$. From this we obtain the following procedure for computations based on this version of iteration:

(1) if $\Delta^{(0)} = x^{(0)} = \beta$, then

$$\Delta^{(k)} = \alpha\Delta^{(k-1)} = \alpha^k\beta \qquad (k = 0, 1, 2, \ldots)$$

and

$$x^{(k)} = \sum_{s=0}^{k} \Delta^{(s)} = \sum_{s=0}^{k} \alpha^s\beta$$

(2) but if $\Delta^{(0)}=x^{(0)}\neq\beta$, then we find

$$\Delta^{(1)}=x^{(1)}-x^{(0)}=\alpha x^{(0)}+\beta-x^{(0)}$$

and assume

$$\Delta^{(k)}=\alpha\Delta^{(k-1)}=\alpha^{k-1}\Delta^{(1)} \qquad (k=1,\ 2,\ 3,\ \ldots)$$

Thus

$$x^{(k)}=\sum_{s=0}^{k}\Delta^{(s)}=x^{(0)}+\sum_{s=1}^{k}\alpha^{s-1}\Delta^{(1)}$$

Example 2. Solve the system

$$\left.\begin{aligned} 2x_1-x_2+x_3&=-3,\\ 3x_1+5x_2-2x_3&=1,\\ x_1-4x_2+10x_3&=0 \end{aligned}\right\} \tag{11}$$

Solution. Reduce the system (11) to the form (2):

$$\begin{aligned} x_1&=-1.5+0.5x_2-0.5x_3,\\ x_2&=0.2-0.6x_1+0.4x_3,\\ x_3&=-0.1x_1+0.4x_2 \end{aligned}$$

Here

$$\alpha=\begin{bmatrix} 0 & 0.5 & -0.5\\ -0.6 & 0 & 0.4\\ -0.1 & 0.4 & 0 \end{bmatrix}$$

and

$$\beta=\begin{bmatrix} -1.5\\ 0.2\\ 0 \end{bmatrix}$$

Using formulas (8) and (9), we get

$$\Delta^{(0)}=\beta=\begin{bmatrix} -1.5\\ 0.2\\ 0 \end{bmatrix},$$

$$\Delta^{(1)}=\alpha\Delta^{(0)}=\begin{bmatrix} 0 & 0.5 & -0.5\\ -0.6 & 0 & 0.4\\ -0.1 & 0.4 & 0 \end{bmatrix}\begin{bmatrix} -1.5\\ 0.2\\ 0 \end{bmatrix}=\begin{bmatrix} 0.1\\ 0.9\\ 0.23 \end{bmatrix},$$

$$\Delta^{(2)}=\alpha\Delta^{(1)}=\begin{bmatrix} 0 & 0.5 & -0.5\\ -0.6 & 0 & 0.4\\ -0.1 & 0.4 & 0 \end{bmatrix}\begin{bmatrix} 0.1\\ 0.9\\ 0.23 \end{bmatrix}=\begin{bmatrix} 0.335\\ 0.032\\ 0.350 \end{bmatrix}$$

and so forth. The results are entered in Table 21.

TABLE 21

SOLVING A LINEAR SYSTEM BY THE MODIFIED ITERATION METHOD (METHOD OF ACCUMULATION)

k	$\Delta^{(k)} x_1$	$\Delta^{(k)} x_2$	$\Delta^{(k)} x_3$
0	—1.500	0.200	0.000
1	0.100	0.900	0.230
2	0.335	0.032	0.350
3	—0.159	—0.061	—0.021
4	—0.020	0.011	—0.008
5	0.010	0.009	0.006
6	0.002	—0.004	0.003
7	—0.004	0.000	—0.001
8	0.000	0.002	0.000
9	0.001	0.000	0.001
Σ	—1.235	1.089	0.560

Thus the approximate values of the roots are

$$x_1 = -1.235, \qquad x_2 = 1.089, \qquad x_3 = 0.560$$

A defect of this version of the method of iteration is the systematic accumulation of errors with increasing number of terms, and, as a result, considerable errors in the required roots. What is more, an error committed in the computations affects the final result. For this reason, the first version of the method of iteration is more reliable.

Remarks concerning computational accuracy. If all the coefficients and constant terms of the given system are exact numbers, the solution by means of the method of successive approximations can be obtained to any preassigned number m of correct decimal places. In this case, retain $m+1$ decimal places in the values of the successive approximations and compute the successive approximations until they coincide. Then round off one digit. If the coefficients and constant terms of the given system are approximate numbers, written to p digits, the solution of the system is carried to $m=p$ digits, as in the case of exact numbers.

We give without proof a sufficient condition for the convergence of the process of iteration (for the proof see Sec. 9.1).

Theorem. *If for the reduced system* (2) *at least one of the following two conditions is valid:*

$$(1) \qquad \sum_{j=1}^{n} |\alpha_{ij}| < 1 \qquad (i=1, 2, \ldots, n)$$

or

$$(2) \qquad \sum_{i=1}^{n} |\alpha_{ij}| < 1 \qquad (j=1, 2, \ldots, n)$$

then the process of iteration (3) *converges to a unique solution of the system irrespective of the choice of the initial approximation.*

Corollary. For the system

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i=1, 2, \ldots, n)$$

the method of iteration converges if the inequalities

$$|a_{ii}| > \sum_{j\neq 1} |a_{ij}| \qquad (i=1, 2, \ldots, n)$$

hold true, that is, if the moduli of the diagonal coefficients are greater for each equation of the system than the sum of the moduli of all the remaining coefficients (disregarding the constant terms).

8.11 REDUCING A LINEAR SYSTEM TO A FORM CONVENIENT FOR ITERATION

The convergence theorem (Sec. 8.10) imposes stringent conditions on the coefficients of the given linear system

$$A\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

However, if $\det A \neq 0$, then by a linear combination of the equations of the system (1), this system can always be replaced by the equivalent system

$$\boldsymbol{x} = \beta + \alpha \boldsymbol{x} \tag{2}$$

such that the conditions of the convergence theorem are valid.

Indeed, multiply (1) by the matrix $D = A^{-1} - \varepsilon$, where $\varepsilon = [\varepsilon_{ij}]$ is a matrix with numerically small elements. Then we have

$$(A^{-1} - \varepsilon) A\boldsymbol{x} = D\boldsymbol{b}$$

or

$$\boldsymbol{x} = \beta + \alpha \boldsymbol{x} \tag{3}$$

where $\alpha = \varepsilon A$ and $\beta = D\boldsymbol{b}$. If $|\varepsilon_{ij}|$ are sufficiently small, it is plain that the system (3) satisfies the conditions of the convergence theorem.

Multiplication by the matrix D is equivalent to a set of elementary transformations of the equations of the system. The problem is to arrive at the standard form (3) with a minimum of effort.

A practical procedure is as follows. Extract from the given system those equations with coefficients whose moduli are greater than the sum of the moduli of the remaining coefficients of the equation. Each chosen equation is written in a row of the new system so that the numerically largest coefficient is a diagonal coefficient.

The remaining unused equations and the chosen equations of the system are made into linearly independent linear combinations in such a way that the above principle of forming a new system is fulfilled and all free rows are filled. Be sure that each equation that was not yet used is involved in at least one linear combination forming an equation of the new system. We give an illustrative example.

Example. Reduce the system

$$\left.\begin{array}{ll}
(A) & 2x_1+3x_2-4x_3+x_4-3=0, \\
(B) & x_1-2x_2-5x_3+x_4-2=0, \\
(C) & 5x_1-3x_2+x_3-4x_4-1=0, \\
(D) & 10x_1+2x_2-x_3+2x_4+4=0
\end{array}\right\}$$

to a form suitable for applying the method of iteration.

Solution. In equation (B) the coefficient of x_3 is greater (in modulus) than the sum of the moduli of the other coefficients, and so we can take this equation for the third equation of the new system. The coefficient of x_1 in equation (D) is also greater than the sum of the moduli of the remaining coefficients of equation (D), and so this equation can be taken for the first equation of the new system. Thus, the new system looks like this:

$$\left.\begin{array}{ll}
(\mathrm{I}) & 10x_1+2x_2-x_3+2x_4+4=0, \\
(\mathrm{II}) & \dots\dots\dots\dots\dots \\
(\mathrm{III}) & x_1-2x_2-5x_3+x_4-2=0, \\
(\mathrm{IV}) & \dots\dots\dots\dots\dots
\end{array}\right\}$$

Examining this system, we can easily see that in order to obtain equation (II) with maximum-modulus coefficient of x_2 it suffices to form the difference $(A)-(C)$:

$$(\mathrm{II}) \qquad x_1+5x_2+x_3+0x_4-1=0$$

The new system now includes the equations (A), (B) and (D), and so equation (IV) must include equation (C) of the given system. A trial convinces us that we can take for equation (IV) the linear combination $2(A)-(B)+2(C)-(D)$:

$$\text{(IV)} \qquad 3x_1+0x_2+0x_3-9x_4-10=0$$

We thus obtain the transformed system of equations I-IV, which is equivalent to the original system and satisfies the conditions of convergence of the iteration process. Solving this system for the diagonal unknowns, we get the system

$$\left.\begin{aligned} x_1 &= \phantom{0.333x_1+{}} 0x_1-0.2x_2+0.1x_3-0.2x_4-0.4,\\ x_2 &= 0.2x_1+0x_2-0.2x_3+0x_4+0.2,\\ x_3 &= 0.2x_1-0.4x_2+0x_3+0.2x_4-0.4,\\ x_4 &= 0.333x_1+0x_2+0x_3+0x_4-1.111 \end{aligned}\right\}$$

to which we can apply the method of iteration.

8.12 THE SEIDEL METHOD

The Seidel method is a certain modification of the method of iteration. The principal idea behind it is that in computing the $(k+1)$th approximation of the unknown x_i, the earlier computed $(k+1)$th approximations of the unknowns $x_1, x_2, \ldots, x_{i-1}$ are taken into account.

Suppose we have a reduced linear system

$$x_i=\beta_i+\sum_{j=1}^{n}\alpha_{ij}x_j \qquad (i=1, 2, \ldots, n)$$

Arbitrarily choose the initial approximations of the roots

$$x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$$

attempting of course to have them roughly correspond to the desired unknowns

$$x_1, x_2, \ldots, x_n$$

Now, assuming that the kth approximations $x_i^{(k)}$ of the roots are known, we, following Seidel, construct the $(k+1)$th approximations of the roots by the following formulas:

$$x_1^{(k+1)}=\beta_1+\sum_{j=1}^{n}\alpha_{1j}x_j^{(k)},$$

$$x_2^{(k+1)}=\beta_2+\alpha_{21}x_1^{(k+1)}+\sum_{j=2}^{n}\alpha_{2j}x_j^{(k)},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_i^{(k+1)}=\beta_i+\sum_{j=1}^{i-1}\alpha_{ij}x_j^{(k+1)}+\sum_{j=i}^{n}\alpha_{ij}x_j^{(k)},$$

. . . .

$$x_n^{(k+1)}=\beta_n+\sum_{j=1}^{n-1}\alpha_{nj}x_j^{(k+1)}+\alpha_{nn}x_n^{(k)} \qquad (k=0,\ 1,\ 2,\ \ldots)$$

Note that the convergence theorem given above (see Sec. 8.10) for simple iteration remains valid for iteration by the Seidel method (see Secs. 9.3 to 9.7).

Ordinarily, the Seidel method yields a better convergence than does the method of simple iteration, what is more, the Seidel process may converge even when the process of iteration diverges. This does not always take place however. Cases are possible when the Seidel process converges more slowly than does the process of iteration. It also sometimes happens that the process of iteration converges while the Seidel process diverges [1] (see Sec. 11.6).

Example. Solve the following system of equations by the Seidel method:

$$\left.\begin{aligned}10x_1+x_2+x_3&=12,\\ 2x_1+10x_2+x_3&=13,\\ 2x_1+2x_2+10x_3&=14\end{aligned}\right\}.$$

Solution. Reduce the system to a form convenient for iteration,

$$\left.\begin{aligned}x_1&=1.2-0.1x_2-0.1x_3,\\ x_2&=1.3-0.2x_1-0.1x_3,\\ x_3&=1.4-0.2x_1-0.2x_2\end{aligned}\right\}$$

For the zeroth approximations of the roots take

$$x_1^{(0)}=1.2,\quad x_2^{(0)}=0,\quad x_3^{(0)}=0$$

Applying the Seidel process, we successively obtain

$$\left.\begin{aligned}x_1^{(1)}&=1.2-0.1\cdot 0-0.1\cdot 0=1.2,\\ x_2^{(1)}&=1.3-0.2\cdot 1.2-0.1\cdot 0=1.06,\\ x_3^{(1)}&=1.4-0.2\cdot 1.2-0.2\cdot 1.06=0.948\end{aligned}\right\}\qquad (\mathrm{I})$$

$$\left.\begin{aligned}x_1^{(2)}&=1.2-0.1\cdot 1.06-0.1\cdot 0.948=0.9992,\\ x_2^{(2)}&=1.3-0.2\cdot 0.9992-0.1\cdot 0.948=1.00536,\\ x_3^{(2)}&=1.4-0.2\cdot 0.9992-0.2\cdot 1.00536=0.999098,\ \text{etc.}\end{aligned}\right\}\qquad (\mathrm{II})$$

The results are computed correct to four decimal places and are tabulated in Table 22.

TABLE 22

FINDING THE ROOTS OF A LINEAR SYSTEM BY THE SEIDEL METHOD

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1.2000	0.0000	0.0000
1	1.2000	1.0600	0.9480
2	0.9992	1.0054	0.9991
3	0.9996	1.0001	1.0001
4	1.0000	1.0000	1.0000
5	1.0000	1.0000	1.0000

The exact values of the roots are $x_1=1$, $x_2=1$, $x_3=1$.

8.13 THE CASE OF A NORMAL SYSTEM

Definition 1. An integral homogeneous polynomial of second degree in n variables is called a *quadratic form* of these variables. In the general case, a quadratic form looks like this

$$u(x_1, x_2, \ldots, x_n)=a_{11}x_1^2+a_{22}x_2^2+\ldots+a_{nn}x_n^2+$$
$$+2a_{12}x_1x_2+2a_{13}x_1x_3+\ldots+2a_{n-1,\,n}x_{n-1}x_n \quad (1)$$

where a_{ij} $(i, j=1, 2, \ldots, n)$ are constants; for the sake of convenience, the coefficients of $i\neq j$ are taken in the even form $2a_{ij}$. Equating u to the constant c, we get the equation of a central quadric surface:

$$u(x_1, x_2, \ldots, x_n)=c$$

in n-dimensional space.

If we put

$$a_{ij}=a_{ji} \quad (2)$$

that is $2a_{ij}=a_{ij}+a_{ji}$, then formula (1) may be written compactly as

$$u(x_1, x_2, \ldots, x_n)=\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \quad (1')$$

The matrix

$$A=[a_{ij}] \quad (3)$$

is called the *matrix of the quadratic form* (1′). By virtue of Condition (2), matrix A will be symmetric, that is, it will coincide with its transpose. Contrariwise, for any symmetric matrix $A=[a_{ij}]$ it is possible to construct an associated quadratic form (1′).

Definition 2. The quadratic form (1) is called *positive (negative) definite* if it assumes positive (negative) values, vanishing only for

$$x_1 = x_2 = \dots = x_n = 0$$

If $u(x_1, x_2, \dots, x_n)$ is a positive definite quadratic form, then the equation

$$u(x_1, x_2, \dots, x_n) = c \qquad (c > 0)$$

is the equation of an ellipsoid. Note that in this case

$$a_{ii} > 0 \qquad (i = 1, 2, \dots, n)$$

since

$$\begin{aligned} a_{11} &= u(1, 0, \dots, 0) > 0, \\ a_{22} &= u(0, 1, \dots, 0) > 0, \\ &\dots\dots\dots\dots \\ a_{nn} &= u(0, 0, \dots, 1) > 0 \end{aligned}$$

Definition 3. Let us call a linear system

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, 2, \dots, n) \tag{4}$$

normal if (1) the matrix $A = [a_{ij}]$ of the coefficients is symmetric, that is, $a_{ij} = a_{ji}$, (2) the corresponding quadratic form $u = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$ is positive definite.

Normal systems are encountered in the solution of many problems, for instance in the method of least squares, when seeking the directions of the principal axes of an ellipsoid, etc.

Reduce the normal system (4), in the ordinary way, to the special form

$$x_i = \sum_{j \neq i} \alpha_{ij} x_j + \beta_i \qquad (i = 1, 2, \dots, n) \tag{4'}$$

where

$$\alpha_{ij} = -\frac{a_{ij}}{a_{ii}} \qquad (j \neq i) \quad \text{and} \quad \beta_i = \frac{b_i}{a_{ii}}$$

Theorem 1. *If the linear system* (4) *is normal, then the Seidel process will always converge for the reduced system* (4') *equivalent to it.*

Proof. See Sec. 11.5 and also [2].

How to reduce a linear system to normal form is indicated by the following theorem.

Theorem 2. *If both members of the linear system*

$$Ax = b \tag{5}$$

with nonsingular matrix $A=[a_{ij}]$ *are premultiplied by the transpose* $A'=[a_{ji}]$, *then the resulting system*

$$A'Ax=A'b \tag{6}$$

is normal.

We first prove that the matrix $A'A$ is symmetric. Indeed, we have

$$(A'A)'=A'A''=A'A$$

We now prove that the quadratic form associated with the matrix $A'A$ is positive definite. We form the quadratic form with matrix $A'A$:

$$u(x_1, x_2, \ldots, x_n)=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ki}a_{kj}x_ix_j$$

Changing the order of summation, we get

$$u=\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ki}x_ia_{kj}x_j=\sum_{k=1}^{n}\left(\sum_{i=1}^{n} a_{ki}x_i\cdot\sum_{j=1}^{n} a_{kj}x_j\right)$$

Since the value of the sum does not depend on the summation index,

$$u=\sum_{k=1}^{n}\left(\sum_{i=1}^{n} a_{ki}x_i\right)^2\geqslant 0$$

By hypothesis, $\det A=\det[a_{ij}]\neq 0$. Therefore the homogeneous system

$$\sum_{i=1}^{n} a_{ki}x_i=0 \qquad (k=1, 2, \ldots, n)$$

has only trivial solutions. Consequently

$$u(x_1, x_2, \ldots, x_n)>0 \text{ for } |x_1|+|x_2|+\ldots+|x_n|\neq 0$$

The proof of the theorem is complete.

8.14 THE METHOD OF RELAXATION

Suppose we have a system of linear equations

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=b_n \end{array}\right\} \tag{1}$$

We transform this system as follows: transpose the constant terms to the left and divide the first equation by $-a_{11}$, the second by $-a_{22}$, etc. We then obtain a system that is *ready for relaxation*:

$$\left.\begin{array}{l} -x_1+b_{12}x_2+\ldots+b_{1n}x_n+c_1=0, \\ b_{21}x_1-\quad x_2+\ldots+b_{2n}x_n+c_2=0, \\ \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ b_{n1}x_1+b_{n2}x_2+\ldots-\quad x_n+c_n=0 \end{array}\right\} \tag{2}$$

where

$$b_{ij}=-\frac{a_{ij}}{a_{ii}}\,(i\neq j) \quad \text{and} \quad c_i=\frac{b_i}{a_{ii}}$$

Let $\boldsymbol{x}^{(0)}=(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ be the initial approximation to the solution of the system (2). Substituting these values into system (2), we get the *residuals*

$$\left.\begin{array}{l} R_1^{(0)}=c_1-x_1^{(0)}+\sum\limits_{j=2}^{n} b_{1j}x_j^{(0)}, \\ R_2^{(0)}=c_2-x_2^{(0)}+\sum\limits_{\substack{j=1\\ j\neq 2}}^{n}{}' b_{2j}x_j^{(0)}, \\ \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ R_n^{(0)}=c_n-x_n^{(0)}+\sum\limits_{j=1}^{n-1} b_{nj}x_j^{(0)} \end{array}\right\} \tag{3}$$

If we give an increment of $\delta x_s^{(0)}$ to one of the unknowns $x_s^{(0)}$, then the corresponding residual $R_s^{(0)}$ will be diminished by the quantity $\delta x_s^{(0)}$, and all the other residuals $R_i^{(0)}\,(i\neq s)$ will be increased by the quantity $b_{is}\delta x_s^{(0)}$. Thus, to make the next residual $R_s^{(1)}$ vanish, it suffices to give the quantity $x_s^{(0)}$ an increment of

$$\delta x_s^{(0)}=R_s^{(0)}$$

and we have

$$R_s^{(1)}=0$$

and

$$R_i^{(1)}=R_i^{(0)}+b_{is}\delta x_s^{(0)} \quad \text{for } i\neq s$$

The *method of relaxation* [3], [4] in its simplest form consists in reducing the numerically largest residual to zero at each step by changing the value of the appropriate component of the approximation. The process is terminated when all the residuals of the last transformed system are equal to zero within the required accuracy. We will not consider the question of the convergence of this process [4].

Example. Solve the following system by the method of relaxation [3]:

$$\left.\begin{aligned} 10x_1 - 2x_2 - 2x_3 &= 6, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ -x_1 - x_2 + 10x_3 &= 8 \end{aligned}\right\} \tag{4}$$

carrying the computations to two decimal places.

Solution. We reduce the system (4) to a form convenient for relaxation:

$$\left.\begin{aligned} -x_1 + 0.2x_2 + 0.2x_3 + 0.6 &= 0, \\ -x_2 + 0.1x_1 + 0.2x_3 + 0.7 &= 0, \\ -x_3 + 0.1x_1 + 0.1x_2 + 0.8 &= 0 \end{aligned}\right\}$$

Choosing the zero values

$$x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$$

for the initial approximations of the roots, we get the respective residuals:

$$R_1^{(0)} = 0.60, \quad R_2^{(0)} = 0.70, \quad R_3^{(0)} = 0.80$$

By the general theory, we assume

$$\delta x_3^{(0)} = 0.80$$

whence we get the residuals

$$\begin{aligned} R_1^{(1)} &= R_1^{(0)} + 0.2 \cdot 0.8 = 0.60 + 0.16 = 0.76, \\ R_2^{(1)} &= R_2^{(0)} + 0.2 \cdot 0.8 = 0.70 + 0.16 = 0.86, \\ R_3^{(1)} &= R_1^{(0)} - R_2^{(0)} = 0 \end{aligned}$$

Now we set

$$\delta x_2^{(1)} = 0.86$$

and so on. The results of the computations are given in Table 23.

Summing all the increments $\delta_i^{(k)}$ ($i = 1, 2, 3$; $k = 0, 1, \ldots$), we get the values of the roots:

$$\begin{aligned} x_1 &= 0 + 0.93 + 0.07 = 1.00, \\ x_2 &= 0 + 0.86 + 0.13 + 0.01 = 1.00, \\ x_3 &= 0 + 0.80 + 0.18 + 0.02 = 1.00 \end{aligned}$$

Check by substituting the values of the roots thus found into the original equations; in this case the system (4) has been solved exactly.

TABLE 23

SOLUTION OF A LINEAR SYSTEM BY THE METHOD OF RELAXATION

	x_1	R_1	x_2	R_2	x_3	R_3
	0	0.60	0	0.70	0	0.80
		0.16		0.16	0.80	—0.80
		0.76		0.86		0
		0.17	0.86	—0.86		0.09
		0.93		0		0.09
	0.93	—0.93		0.09		0.09
		0		0.09		0.18
		0.04		0.04	0.18	—0.18
		0.04		0.13		0
		0.03	0.13	—0.13		0.01
		0.07		0		0.01
	0.07	—0.07		0.01		0.01
		0		0.01		0.02
		0		0	0.02	—0.02
		0		0.01		0
		0	0.01	—0.01		0
		0		0		0
$\sum$	1.00		1.00		1.00	

8.15 CORRECTING ELEMENTS OF AN APPROXIMATE INVERSE MATRIX

Suppose we have a nonsingular matrix A and it is required to find the inverse A^{-1}. Also suppose we have found an approximate value of the inverse matrix $D_0 \approx A^{-1}$. It is then possible to improve the accuracy by using the method of successive approximations in a special form. For a preliminary measure of the error, we use the difference

$$F_0 = E - AD_0$$

If $F_0 = 0$, then plainly $D_0 = A^{-1}$, and so if the moduli of the elements of matrix F_0 are small, then matrices A^{-1} and D_0 are nearly equal. We will construct successive approximations by the formula

$$D_k = D_{k-1} + D_{k-1}F_{k-1} \qquad (k = 1, 2, 3, \ldots) \tag{1}$$

the corresponding error being

$$F_k = E - AD_k$$

Let us estimate the rapidity of convergence of the successive approximations. We have

$$F_1 = E - AD_1 = E - A(D_0 + D_0F_0) = E - AD_0(E + F_0) =$$
$$= E - (E - F_0)(E + F_0) = E - (E - F_0^2) = F_0^2$$

Similarly

$$F_2 = F_1^2 = F_0^4$$

and, generally,

$$F_k = F_0^{2^k} \qquad (k = 1, 2, 3, \ldots) \tag{2}$$

We will prove that if

$$\| F_0 \| \leqslant q < 1 \tag{3}$$

where $\| F_0 \|$ is some canonical norm of the matrix F_0 (Sec. 7.7), then the process of iteration (1) converges, that is,

$$\lim_{k \to \infty} D_k = A^{-1}$$

Indeed, from formula (2) we have

$$\| F_k \| \leqslant \| F_0 \|^{2^k} \leqslant q^{2^k}$$

And so

$$\lim_{k \to \infty} \| F_k \| = 0$$

and, consequently,

$$\lim_{k \to \infty} F_k = \lim_{k \to \infty} (E - AD_k) = 0$$

or

$$E - A \lim_{k \to \infty} D_k = 0$$

that is

$$\lim_{k \to \infty} D_k = A^{-1}E = A^{-1}$$

Thus, the assertion is proved.

In particular, using the m-norm (Sec. 7.7), we find that if the elements of the matrix $F_0 = [f_{ij}]$ satisfy the inequality

$$| f_{ij} | \leqslant \frac{q}{n}$$

where n is the order of the matrix and $0 \leqslant q < 1$, then the process of iteration (1) definitely converges.

Assuming inequality (3) to be valid, we estimate the error

$$R_k = \| A^{-1} - D_k \| \leqslant \| A^{-1} \| \, \| E - AD_k \| = \| A^{-1} \| \, \| F_k \| \leqslant \| A^{-1} \| \, q^{2^k}$$

Since

$$AD_0 = E - F_0$$

it follows that

$$A^{-1} = D_0 (E - F_0)^{-1} = D_0 (E + F_0 + F_0^2 + \ldots)$$

whence

$$\| A^{-1} \| \leqslant \| D_0 \| \{ \| E \| + q + q^2 + \ldots \} = \| D_0 \| \left\{ \| E \| + \frac{q}{1-q} \right\}$$

For the m-norm or the l-norm we have $\| E \| = 1$, and so

$$\| A^{-1} \| < \frac{\| D_0 \|}{1-q}$$

Thus

$$\| A^{-1} - D_k \| \leqslant \frac{\| D_0 \|}{1-q} \| F_k \| \tag{4}$$

or

$$\| A^{-1} - D_k \| \leqslant \frac{\| D_0 \|}{1-q} q^{2^k} \tag{5}$$

where the norm is to be understood in the sense of the m-norm or l-norm. From formula (4) it follows that the convergence of the process (1) is very rapid for $q \ll 1$.

In practical situations, the process of improving the elements of the inverse matrix is terminated when the inequality

$$\| D_k - D_{k-1} \| \leqslant \varepsilon$$

where ε is the specified accuracy, is ensured.

Example. Correct the elements of the approximate inverse matrix obtained in the example of Sec. 8.7.

Solution. Using the Gaussian method, we obtain for the matrix

$$A = \begin{bmatrix} 1.8 & -3.8 & 0.7 & -3.7 \\ 0.7 & 2.1 & -2.6 & -2.8 \\ 7.3 & 8.1 & 1.7 & -4.9 \\ 1.9 & -4.3 & -4.9 & -4.7 \end{bmatrix}$$

the approximate inverse

$$D_0 = \begin{bmatrix} -0.21121 & -0.46003 & 0.16284 & 0.26956 \\ -0.03533 & 0.16873 & 0.01573 & -0.08920 \\ 0.23030 & 0.04607 & -0.00944 & -0.19885 \\ -0.29316 & -0.38837 & -0.06128 & 0.18513 \end{bmatrix}$$

Here

$$AD_0 = E - 10^{-3}\cdot\begin{bmatrix} 0.03 & 0.00 & 0.01 & 0.00 \\ 0.25 & 0.03 & 0.02 & 0.39 \\ 8.08 & 10.17 & 0.18 & -0.09 \\ 0.00 & 0.00 & 0.00 & -0.48 \end{bmatrix}$$

whence

$$F_0 = E - AD_0 = 10^{-3}\cdot\begin{bmatrix} 0.03 & 0.00 & 0.01 & 0.00 \\ 0.25 & 0.03 & 0.02 & 0.39 \\ 8.08 & 10.17 & 0.18 & -0.09 \\ 0.00 & 0.00 & 0.00 & -0.48 \end{bmatrix}$$

To further improve the elements of the matrix D_0 let us use the iteration process

$$D_{k+1} = D_k + D_k F_k, \quad F_k = E - AD_k \qquad (k = 0, 1, 2, \ldots)$$

Since

$$q = \|F_0\|_l = 10^{-3}\cdot(0.03 + 10.17) = 1.02\cdot 10^{-2} \leqslant 1$$

the iteration process converges rapidly.

We have

$$D_0F_0 = \begin{bmatrix} -0.21121 & -0.46003 & 0.16284 & 0.26956 \\ -0.03533 & 0.16873 & 0.01573 & -0.08920 \\ 0.23030 & 0.04607 & -0.00944 & -0.19885 \\ -0.29316 & -0.38837 & -0.06128 & 0.18513 \end{bmatrix} \times$$

$$\times 10^{-3}\cdot\begin{bmatrix} 0.03 & 0.00 & 0.01 & 0.00 \\ 0.25 & 0.03 & 0.02 & 0.39 \\ 8.08 & 10.17 & 0.18 & -0.09 \\ 0.00 & 0.00 & 0.00 & -0.48 \end{bmatrix} =$$

$$= 10^{-3}\cdot\begin{bmatrix} 1.19 & 1.64 & 0.02 & -0.32 \\ 0.17 & 0.16 & 0.01 & 0.11 \\ -0.06 & -0.09 & 0.00 & 0.11 \\ 0.39 & 0.61 & 0.00 & -0.24 \end{bmatrix}$$

whence

$$D_1 = D_0 + D_0F_0 = \begin{bmatrix} -0.21121 & -0.46003 & 0.16284 & 0.26956 \\ -0.03533 & 0.16873 & 0.01573 & -0.08920 \\ 0.23030 & 0.04607 & -0.00944 & -0.19885 \\ -0.29316 & -0.38837 & -0.06128 & 0.18513 \end{bmatrix} +$$

$$+10^{-3}\cdot\begin{bmatrix} 1.19 & 1.64 & 0.02 & -0.32 \\ 0.17 & 0.16 & 0.01 & 0.11 \\ -0.06 & -0.09 & 0.00 & 0.11 \\ 0.39 & 0.61 & 0.00 & -0.24 \end{bmatrix}=$$

$$=\begin{bmatrix} -0.21002 & -0.45839 & 0.16286 & 0.26924 \\ -0.03516 & 0.16889 & 0.01574 & -0.08909 \\ 0.23024 & 0.04598 & -0.00944 & -0.19874 \\ -0.29277 & -0.38776 & -0.06128 & 0.18489 \end{bmatrix}$$

We can take it that

$$A^{-1}\approx D_1$$

since

$$AD_1 \doteq E-10^{-5}\cdot\begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and

$$F_1=E-AD_1=10^{-5}\cdot\begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

For the error we have the estimate, on the basis of formula (4),

$$\| A^{-1}-D_1 \|_l \leqslant \frac{\| D_0 \|_l}{1-q} \| F_1 \|_l$$

Since

$$\| D_0 \|_l = 0.46003+0.16873+0.04607+0.38837 < 1.07$$

and

$$\| F_1 \|_l = 10^{-5}\cdot(2+2+4)=8\cdot 10^{-5}$$

we finally get

$$\| A^{-1}-D_1 \|_l \leqslant \frac{1.07}{1-1.02\cdot 10^{-2}}\cdot 8\cdot 10^{-5} < 9\cdot 10^{-5}$$

Note. Choosing the approximate inverse matrix can be done in a variety of ways. In particular, use is made of the method of matrix inversion given in Sec. 7.12.

We conclude this chapter with the remark that many other methods for solving systems of linear algebraic equations have been elaborated (the method of Purcell, the escalator method [6], the method of Richardson [7] and others).

REFERENCES FOR CHAPTER 8

[1] *V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1950, Chapter II (in Russian).

[2] *James B. Scarborough, Numerical Mathematical Analysis*, 1955, Chapter XVIII.

[3] *Mario G. Salvadori and M. L. Baron, Numerical Methods in Engineering*, 1952, Chapter I, Sec. 10.

[4] *Edwin F. Beckenbach (editor), Modern Mathematics for the Engineer*, 1956, First Series, Chapter 17, What Are Relaxation Methods? by George E. Forsythe.

[5] *Kh. L. Smolitsky, Computational Mathematics* (Lecture Notes), 1960 (in Russian).

[6] *D. K. Faddeyev and V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1960, Chapter II (in Russian).

[7] *I. S. Berezin and N. P. Zhidkov, Computational Methods*, 1959, Chapter VI (in Russian).

*Chapter 9

THE CONVERGENCE OF ITERATION PROCESSES FOR SYSTEMS OF LINEAR EQUATIONS

9.1 SUFFICIENT CONDITIONS FOR THE CONVERGENCE OF THE ITERATION PROCESS

Supposé we have a reduced linear system:

$$\boldsymbol{x} = \alpha \boldsymbol{x} + \beta \tag{1}$$

where

$$\alpha = [\alpha_{ij}], \quad \beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix}$$

are a given matrix and a given vector and $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$ is the desired vector.

Theorem. *For the reduced linear system* (1) *the process of iteration converges to its unique solution if some canonical norm of the matrix* α *is less than unity; that is, a sufficient condition for convergence of the iteration process*

$$\boldsymbol{x}^{(k)} = \beta + \alpha \boldsymbol{x}^{(k-1)} \qquad (k = 1, 2, \ldots)$$

($\mathrm{x}^{(0)}$ *arbitrary) is*

$$\|\alpha\| < 1 \tag{2}$$

Proof. Starting with an arbitrary vector $\boldsymbol{x}^{(0)}$, we construct a sequence of approximations

$$\begin{aligned} \boldsymbol{x}^{(1)} &= \beta + \alpha \boldsymbol{x}^{(0)}, \\ \boldsymbol{x}^{(2)} &= \beta + \alpha \boldsymbol{x}^{(1)}, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \boldsymbol{x}^{(k)} &= \beta + \alpha \boldsymbol{x}^{(k-1)} \end{aligned}$$

whence

$$\boldsymbol{x}^{(k)} = (E + \alpha + \alpha^2 + \ldots + \alpha^{k-1}) \beta + \alpha^k \boldsymbol{x}^{(0)} \tag{3}$$

Since for $\|\alpha\|<1$ we have $\|\alpha^k\|\to 0$ as $k\to\infty$, it follows (see Sec. 7.10) that

$$\lim_{k\to\infty}\alpha^k=0$$

and

$$\lim_{k\to\infty}(E+\alpha+\alpha^2+\ldots+\alpha^{k-1})=\sum_{k=0}^{\infty}\alpha^k=(E-\alpha)^{-1}$$

And so, passing to the limit in (3) as $k\to\infty$, we get

$$\boldsymbol{x}=\lim_{k\to\infty}\boldsymbol{x}^{(k)}=(E-\alpha)^{-1}\beta \tag{4}$$

This proves the convergence of the iterative process. Moreover, from (4) we have

$$(E-\alpha)\boldsymbol{x}=\beta$$

or

$$\boldsymbol{x}=\alpha\boldsymbol{x}+\beta$$

which means that the limiting vector $\boldsymbol{x}$ is a solution of system (1). Since the matrix $E-\alpha$ of system (1) is nonsingular, the solution $\boldsymbol{x}$ is unique.

Corollary 1. The iteration process for system (1) converges if

(a) $\|\alpha\|_m=\max\limits_i\sum\limits_{j=1}^{n}|\alpha_{ij}|<1$

or

(b) $\|\alpha\|_l=\max\limits_j\sum\limits_{i=1}^{n}|\alpha_{ij}|<1$

or

(c) $\|\alpha\|_k=\sqrt{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}|\alpha_{ij}|^2}<1$

In particular, the iteration process definitely converges if the elements of matrix α satisfy the inequality

$$|\alpha_{ij}|<\frac{1}{n}$$

where n is the number of unknowns in system (1).

Indeed, (a), (b) and (c) are the simplest canonical norms of the matrix α.

Corollary 2. For the system

$$\sum_{j=1}^{n}a_{ij}x_j=b_i\qquad(i=1,\,2,\,\ldots,\,n) \tag{5}$$

the process of iteration converges if the following inequalities hold:

$$\text{(a}'\text{)} \qquad |a_{ii}| > \sum_{j=1}^{n}{}' |a_{ij}| \qquad (i=1, 2, \ldots, n)$$

or

$$\text{(b}'\text{)} \qquad |a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \qquad (j=1, 2, \ldots, n)$$

where the prime on the summation symbol means that the values $i=j$ are dropped in the summation; that is, convergence occurs if the moduli of the diagonal elements of the matrix $A=[a_{ij}]$ of system (1) either, for each row, exceed the sum of the moduli of the nondiagonal elements of that row or, for each column, exceed the sum of the moduli of the nondiagonal elements of that column.

Indeed, given inequality (a′), the respective inequality (a) of Corollary 1 will clearly hold.

To prove the second assertion, in (5) put

$$x_i = \frac{z_i}{a_{ii}} \qquad (i=1, 2, \ldots, n)$$

where z_i are the new unknowns. We then get the system

$$\sum_{j=1}^{n} \frac{a_{ij}}{a_{jj}} z_j = b_i \qquad (i=1, 2, \ldots, n) \tag{5′}$$

for which the iteration process either converges or diverges simultaneously with the process of iteration of the original system (5). Reducing (5′) to the special form (1) in the ordinary way, and utilizing Condition (b) of Corollary 1, we get a sufficient condition for the convergence of the process of iteration of the system (5):

$$\sum_{i=1}^{n}{}' \left| \frac{a_{ij}}{a_{jj}} \right| < 1 \qquad (j=1, 2, \ldots, n)$$

or

$$|a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \qquad (j=1, 2, \ldots, n)$$

9.2 AN ESTIMATE OF THE ERROR OF APPROXIMATIONS IN THE ITERATION PROCESS

Let $\boldsymbol{x}^{(k-1)}$ and $\boldsymbol{x}^{(k)}$ $(k \geqslant 1)$ be two successive approximations of the solution of the linear system $\boldsymbol{x} = \alpha \boldsymbol{x} + \beta$. For $p \geqslant 1$, we have

$$\begin{aligned} \| \boldsymbol{x}^{(k+p)} - \boldsymbol{x}^{(k)} \| \leqslant \| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \| + \| \boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)} \| + \\ + \ldots + \| \boldsymbol{x}^{(k+p)} - \boldsymbol{x}^{(k+p-1)} \| \end{aligned} \tag{1}$$

Since

$$x^{(m+1)} = \alpha x^{(m)} + \beta$$

and

$$x^{(m)} = \alpha x^{(m-1)} + \beta$$

it follows that

$$x^{(m+1)} - x^{(m)} = \alpha (x^{(m)} - x^{(m-1)})$$

and, hence,

$$\| x^{(m+1)} - x^{(m)} \| \leqslant \| \alpha \| \, \| x^{(m)} - x^{(m-1)} \| \leqslant$$
$$\leqslant \| \alpha \|^{m-k} \| x^{(k+1)} - x^{(k)} \| \quad \text{for} \quad m > k \geqslant 1$$

Therefore, from formula (1) we get

$$\| x^{(p+k)} - x^{(k)} \| \leqslant \| x^{(k+1)} - x^{(k)} \| + \| \alpha \| \, \| x^{(k+1)} - x^{(k)} \| + \ldots +$$
$$+ \| \alpha \|^{p-1} \| x^{(k+1)} - x^{(k)} \| \leqslant \frac{1}{1 - \| \alpha \|} \| x^{(k+1)} - x^{(k)} \|$$

Passing to the limit in the last inequality as $p \to \infty$, we finally obtain

$$\| x - x^{(k)} \| \leqslant \frac{\| x^{(k+1)} - x^{(k)} \|}{1 - \| \alpha \|} \tag{2}$$

for $k \geqslant 1$, or

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|}{1 - \| \alpha \|} \| x^{(k)} - x^{(k-1)} \|$$

If

$$\| \alpha \| \leqslant \frac{1}{2}$$

then the preceding formula becomes

$$\| x - x^{(k)} \| \leqslant \| x^{(k)} - x^{(k-1)} \|$$

Thus, in this case, if

$$\| x^{(k)} - x^{(k-1)} \| < \varepsilon$$

then we also have

$$\| x - x^{(k)} \| < \varepsilon$$

In the general case, if in the process of computations we find that

$$\| x^{(k)} - x^{(k-1)} \| \leqslant \frac{1-q}{q} \varepsilon$$

where $q = \| \alpha \| < 1$, then

$$\| x - x^{(k)} \| \leqslant \varepsilon$$

and thus

$$| x_i - x_i^{(k)} | \leqslant \varepsilon \qquad (i = 1, 2, \ldots, n)$$

It is of course assumed here that the successive approximations $\boldsymbol{x}^{(j)}$ ($j=0, 1, \ldots, k$) are computed exactly, which is to say that rounding errors are completely absent.

Utilizing the above obtained estimates for the norm of the difference of two successive approximations, we get, from formula (2),

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\| \leqslant \frac{\|\alpha\|^k}{1-\|\alpha\|} \|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|$$

In particular, if we choose

$$\boldsymbol{x}^{(0)}=\beta$$

then

$$\boldsymbol{x}^{(1)}=\alpha\beta+\beta$$

and

$$\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\| = \|\alpha\beta\| \leqslant \|\alpha\| \|\beta\|$$

Hence

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\| \leqslant \frac{\|\alpha\|^{k+1}}{1-\|\alpha\|} \|\beta\| \tag{2'}$$

Example. Show that the process of iteration converges for the following system:

$$\left.\begin{aligned} 10x_1 - x_2 + 2x_3 - 3x_4 &= 0, \\ x_1 + 10x_2 - x_3 + 2x_4 &= 5, \\ 2x_1 + 3x_2 + 20x_3 - x_4 &= -10, \\ 3x_1 + 2x_2 + x_3 + 20x_4 &= 15 \end{aligned}\right\} \tag{3}$$

How many iterations have to be carried out to find the roots of system (3) to within 10^{-4}?

Solution. Reducing (3) to the special form, we get

$$\left.\begin{aligned} x_1 &= 0.1\,x_2 - 0.2\,x_3 + 0.3x_4, \\ x_2 &= -0.1\,x_1 + 0.1\,x_3 - 0.2x_4 + 0.5, \\ x_3 &= -0.1\,x_1 - 0.15x_2 + 0.05x_4 - 0.5, \\ x_4 &= -0.15x_1 - 0.1\,x_2 - 0.05x_3 + 0.75 \end{aligned}\right\} \tag{3'}$$

Then the matrix of the system is

$$\alpha = \begin{bmatrix} 0 & 0.1 & -0.2 & 0.3 \\ -0.1 & 0 & 0.1 & -0.2 \\ -0.1 & -0.15 & 0 & 0.05 \\ -0.15 & -0.1 & -0.05 & 0 \end{bmatrix}$$

Using, say, the norm $\|\alpha\|_l$, we get

$$\|\alpha\|_l = \max(0.35,\ 0.35,\ 0.35,\ 0.55) = 0.55 < 1$$

Hence the process of iteration of the system (3′) converges.

For the initial approximation of the root $\boldsymbol{x}$ we take

$$\boldsymbol{x}^{(0)}=\beta=\begin{bmatrix} 0 \\ 0.5 \\ -0.5 \\ 0.75 \end{bmatrix}$$

whence

$$\|\beta\|_l = 0+0.5+0.5+0.75=1.75$$

Let k be the number of iterations required to achieve the specified accuracy. Using formula (2′), we have

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\| \leqslant \frac{\|\alpha\|_l^{k+1}\,\|\beta\|_l}{1-\|\alpha\|_l}=\frac{0.55^{k+1}\cdot 1.75}{0.45}<10^{-4}$$

From this,

$$0.55^{k+1}<\frac{45}{175}\cdot 10^{-4}$$

and

$$(k+1)\log_{10}0.55<\log_{10}45-\log_{10}175-4$$

or

$$-(k+1)\cdot 0.25964<1.65321-2.24304-4=-4.58983$$

Consequently

$$k+1>\frac{4.58983}{0.25964}\approx 17.7$$

and

$$k>16.7$$

We can take $k=17$.

It should be pointed out that the theoretical estimate of the number of iterations necessary to ensure the specified accuracy turns out to be excessively high.

9.3 FIRST SUFFICIENT CONDITION FOR CONVERGENCE OF THE SEIDEL PROCESS

Theorem. *If for a linear system*

$$\boldsymbol{x}=\alpha\boldsymbol{x}+\beta \tag{1}$$

the condition

$$\|\alpha\|_m<1 \tag{2}$$

where

$$\|\alpha\|_m=\max_i \sum_{j=1}^{n}|\alpha_{ij}|$$

is fulfilled, then the Seidel process converges for (1) *to its unique solution for any choice of the initial vector* $\mathbf{x}^{(0)}$.

Proof. Let $\boldsymbol{x}^{(k)}=\{x_1^{(k)}, \ldots, x_n^{(k)}\}$ be the kth approximation in the Seidel process. We then have

$$x_i^{(k)}=\sum_{j=1}^{i-1} \alpha_{ij}x_j^{(k)}+\sum_{j=i}^{n} \alpha_{ij}x_j^{(k-1)}+\beta_i \tag{3}$$
$$(i=1, 2, \ldots, n; \quad k=1, 2, \ldots)$$

If Condition (2) is fulfilled, the system (1) admits the unique solution $\boldsymbol{x}=\{x_1, \ldots, x_n\}$, which may be found, say, by the method of iteration. We have

$$x_i=\sum_{j=1}^{n} \alpha_{ij}x_j+\beta_i \qquad (i=1, 2, \ldots) \tag{4}$$

Subtracting (3) from (4), we get

$$x_i-x_i^{(k)}=\sum_{j=1}^{i-1} \alpha_{ij}(x_j-x_j^{(k)})+\sum_{j=i}^{n}\alpha_{ij}(x_j-x_j^{(k-1)})$$

whence

$$|x_i-x_i^{(k)}|\leqslant\sum_{j=1}^{i-1} |\alpha_{ij}||x_j-x_j^{(k)}|+\sum_{j=i}^{n} |\alpha_{ij}||x_j-x_j^{(k-1)}| \tag{5}$$
$$(i=1, 2, \ldots, n)$$

According to the meaning of the accepted norm,

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m=\max_i |x_i-x_i^{(k)}|$$

and so

$$|x_j-x_j^{(k)}|\leqslant\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m$$

$(j=1, 2, \ldots, n)$. Hence, from inequality (5) we derive

$$|x_i-x_i^{(k)}|\leqslant p_i\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m+q_i\|\boldsymbol{x}-\boldsymbol{x}^{(k-1)}\|_m \tag{6}$$

where

$$p_i=\sum_{j=1}^{i-1} |\alpha_{ij}| \quad \text{and} \quad q_i=\sum_{j=i}^{n} |\alpha_{ij}|$$

Let $s=s(k)$ be the value of the index i for which

$$|x_s-x_s^{(k)}|=\max_i |x_i-x_i^{(k)}|=\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m$$

Assuming $i=s$ in inequality (6), we get

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m\leqslant p_s\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m+q_s\|\boldsymbol{x}-\boldsymbol{x}^{(k-1)}\|_m$$

or

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m\leqslant\frac{q_s}{1-p_s}\|\boldsymbol{x}-\boldsymbol{x}^{(k-1)}\|_m$$

whence

$$\| x - x^{(k)} \|_m \leqslant \mu \| x - x^{(k-1)} \|_m \tag{7}$$

where

$$\mu = \max_i \frac{q_i}{1 - p_i} \tag{8}$$

We will show that

$$\mu \leqslant \| \alpha \|_m < 1$$

Indeed, since

$$p_i + q_i = \sum_{j=1}^{n} | \alpha_{ij} | \leqslant \| \alpha \|_m < 1$$

it follows that

$$q_i \leqslant \| \alpha \|_m - p_i$$

and hence

$$\frac{q_i}{1 - p_i} \leqslant \frac{\| \alpha \|_m - p_i}{1 - p_i} \leqslant \frac{\| \alpha \|_m - p_i \| \alpha \|_m}{1 - p_i} = \| \alpha \|_m$$

Therefore

$$\mu = \| \alpha \|_m < 1$$

From inequality (7) it follows that

$$\| x - x^{(k)} \|_m \leqslant \mu^k \| x - x^{(0)} \|_m$$

and consequently

$$\lim_{k \to \infty} x^{(k)} = x$$

This completes the proof that the Seidel process converges to the required solution.

Note. Since for the method of iteration we have

$$\| x - x^{(k)} \| \leqslant \| \alpha \|_m \| x - x^{(k-1)} \|$$

and for the Seidel method we obtain

$$\| x - x^{(k)} \| \leqslant \mu \| x - x^{(k-1)} \|$$

where $\mu \leqslant \| \alpha \|_m$, it follows that under the conditions of the theorem the convergence of the Seidel process is in general somewhat better than the convergence of the process of simple iteration. From formula (8) it follows that in this case when we use the Seidel method, it is convenient to arrange the system (1) so that the first equation of the system has the smallest sum of the moduli of the coefficients:

$$q_1 = \sum_{j=1}^{n} | \alpha_{1j} |$$

9.4 ESTIMATING THE ERROR OF APPROXIMATIONS IN THE SEIDEL PROCESS BY THE m-NORM

Let $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^{(k+1)}$ be two successive iterations in the Seidel process. Applying to these iterations the transformations utilized in the proof of the theorem of Sec. 9.3, we get an inequality similar to the inequality (7) of Sec. 9.3:

$$\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_m \leqslant \mu \|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m$$

From this

$$\begin{aligned}\|\boldsymbol{x}^{(k+p)}-\boldsymbol{x}^{(k)}\|_m \leqslant{}& \|\boldsymbol{x}^{(k+p)}-\boldsymbol{x}^{(k+p-1)}\|_m + \\ &+\|\boldsymbol{x}^{(k+p-1)}-\boldsymbol{x}^{(k+p-2)}\|_m+\ldots+\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_m \leqslant \\ &\leqslant \mu^p\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m+\mu^{p-1}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m+\ldots \\ &\ldots+\mu\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m \leqslant \frac{\mu}{1-\mu}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m\end{aligned}$$

As $p \to \infty$ we get

$$\lim_{p\to\infty} \boldsymbol{x}^{(k+p)} = \boldsymbol{x}$$

and hence

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m \leqslant \frac{\mu}{1-\mu}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|_m$$

where

$$\mu = \max_i \frac{\sum_{j=i}^{n}|\alpha_{ij}|}{1-\sum_{j=1}^{i-1}|\alpha_{ij}|} \leqslant \|\alpha\|_m$$

In particular, we derive

$$\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_m \leqslant \frac{\mu^k}{1-\mu}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|_m$$

from the inequality obtained, that is,

$$|x_i-x_i^{(k)}| \leqslant \frac{\mu^k}{1-\mu}\max_j |x_j^{(1)}-x_j^{(0)}| \qquad (i=1,\ 2,\ \ldots,\ n)$$

9.5 SECOND SUFFICIENT CONDITION FOR CONVERGENCE OF THE SEIDEL PROCESS

Theorem. *If for a linear system*

$$\boldsymbol{x} = \alpha\boldsymbol{x}+\beta \tag{1}$$

the condition

$$\|\alpha\|_l < 1$$

where

$$\|\alpha\|_l = \max_j \sum_{i=1}^{n} |\alpha_{ij}|$$

is fulfilled, then the Seidel process converges to a unique solution of system (1) *for any choice of the initial vector.*

Proof. Suppose

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k-1)} + \beta_i \qquad \begin{matrix} (i=1,\ 2,\ \ldots,\ n; \\ k=1,\ 2,\ \ldots) \end{matrix} \tag{2}$$

For the exact solution $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, which exists and is unique, we have

$$x_i = \sum_{j=1}^{i-1} \alpha_{ij} x_j + \sum_{j=i}^{n} \alpha_{ij} x_j + \beta_i \tag{3}$$

Subtracting from (3) the corresponding equations (2), we get

$$x_i - x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} (x_j - x_j^{(k)}) + \sum_{j=i}^{n} \alpha_{ij} (x_j - x_j^{(k-1)})$$

whence

$$|x_i - x_i^{(k)}| \leqslant \sum_{j=1}^{i-1} |\alpha_{ij}| |x_j - x_j^{(k)}| + \sum_{j=i}^{n} |\alpha_{ij}| |x_j - x_j^{(k-1)}|$$
$$(i = 1,\ 2,\ \ldots,\ n)$$

Summing these inequalities, we get

$$\sum_{i=1}^{n} |x_i - x_1^{(k)}| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{i-1} |\alpha_{ij}| |x_j - x_j^{(k)}| + \sum_{i=1}^{n} \sum_{j=i}^{n} |\alpha_{ij}| |x_j - x_j^{(k-1)}|$$

or, changing the order of the summation, we obtain

$$\sum_{i=1}^{n} |x_i - x_i^{(k)}| \leqslant \sum_{j=1}^{n-1} |x_j - x_j^{(k)}| \sum_{i=j+1}^{n} |\alpha_{ij}| + \sum_{j=1}^{n} |x_j - x_j^{(k-1)}| \sum_{i=1}^{j} |\alpha_{ij}| \tag{4}$$

Set

$$s_j = \sum_{i=j+1}^{n} |\alpha_{ij}|, \quad t_j = \sum_{i=1}^{j} |\alpha_{ij}| \qquad (j = 1,\ 2,\ \ldots,\ n-1)$$

and

$$s_n = 0, \quad t_n = \sum_{i=1}^{n} |\alpha_{ij}|$$

Obviously

$$s_j + t_j = \sum_{i=1}^{n} |\alpha_{ij}| \leqslant \|\alpha\|_l < 1$$

whence

$$s_j < 1$$

Inequality (4) takes the form

$$\sum_{i=1}^{n} |x_i - x_i^{(k)}| \leqslant \sum_{j=1}^{n} s_j |x_j - x_j^{(k)}| + \sum_{j=1}^{n} t_j |x_j - x_j^{(k-1)}|$$

or

$$\sum_{j=1}^{n} (1 - s_j) |x_j - x_j^{(k)}| \leqslant \sum_{j=1}^{n} t_j |x_j - x_j^{(k-1)}|$$

Since

$$t_j \leqslant \|\alpha\|_l - s_j \leqslant \|\alpha\|_l - s_j \|\alpha\|_l = \|\alpha\|_l (1 - s_j) \tag{5}$$

we then have

$$\sum_{j=1}^{n} (1 - s_j) |x_j - x_j^{(k)}| \leqslant \|\alpha\|_l \sum_{i=1}^{n} (1 - s_j) |x_j - x_j^{(k-1)}| \leqslant$$

$$\leqslant \|\alpha\|_l^k \sum_{i=1}^{n} (1 - s_j) |x_j - x_j^{(0)}| \tag{6}$$

Whence, passing to the limit as $k \to \infty$ and noting that $\|\alpha\|_l < 1$, we get

$$\lim_{k \to \infty} \sum_{j=1}^{n} (1 - s_j) |x_j - x_j^{(k)}| = 0$$

Hence

$$\lim_{k \to \infty} x_j^{(k)} = x_j \qquad (j = 1, 2, \ldots, n)$$

and the proof is complete.

9.6 ESTIMATING THE ERROR OF APPROXIMATIONS IN THE SEIDEL PROCESS BY THE *l*-NORM

Suppose

$$\sigma_{k+1} = \sum_{j=1}^{n} (1 - s_j) |x_j^{(k+1)} - x_j^{(k)}| \qquad (k = 0, 1, 2, \ldots)$$

Utilizing transformations similar to those used in the proof of the theorem of the preceding section, we get the following inequality [inequality (6) of Sec. 9.5] for two successive iterations $x_j^{(k)}$ and $x_j^{(k+1)}$:

$$\sigma_{k+1} \leqslant \rho \sigma_k \tag{1}$$

where, by virtue of inequality 5 of Sec. 9.5,

$$\rho = \max_j \frac{t_j}{1 - s_j} \leqslant \|\alpha\|_l$$

whence

$$\sigma_{k+p} \leqslant \rho^p \sigma_k \qquad (p=1, 2, \ldots)$$

We then have

$$\sum_{j=1}^{n} (1-s_j) \, | x_j^{(k+p)} - x_j^{(k)} | \leqslant \sigma_{k+p} + \sigma_{k+p-1} + \ldots + \sigma_{k+1} \leqslant$$
$$\leqslant \rho^p \sigma_k + \rho^{p-1} \sigma_k + \ldots + \rho \sigma_k \leqslant \frac{\rho \sigma_k}{1-\rho}$$

From this, we get, as $p \to \infty$,

$$\sum_{j=1}^{n} (1-s_j) \, | x_j - x_j^{(k)} | \leqslant \frac{\rho \sigma_k}{1-\rho}$$

or

$$\sum_{j=1}^{n} | x_j - x_j^{(k)} | \leqslant \frac{\rho}{(1-s)(1-\rho)} \sum_{j=1}^{n} | x_j^{(k)} - x_j^{(k-1)} |$$

where

$$s = \max_j s_j = \max_j \sum_{i=j+1}^{n} |\alpha_{ij}|$$

Since it follows from formula (1) that

$$\sigma_k \leqslant \rho^{k-1} \sigma_1$$

the estimate

$$\| \boldsymbol{x}_j - \boldsymbol{x}_j^{(k)} \|_l = \sum_{j=1}^{n} | x_j - x_j^{(k)} | \leqslant \frac{\rho^k}{(1-s)(1-\rho)} \sigma_1 \leqslant$$
$$\leqslant \frac{\rho^k}{(1-s)(1-\rho)} \sum_{j=1}^{n} | x_j^{(1)} - x_j^{(0)} |$$

is also valid.

9.7 THIRD SUFFICIENT CONDITION FOR CONVERGENCE OF THE SEIDEL PROCESS

Theorem. *If for a linear system*

$$\boldsymbol{x} = \alpha \boldsymbol{x} + \beta \tag{1}$$

the condition

$$\| \alpha \|_k < 1$$

where

$$\| \alpha \|_k = \sqrt{\sum_{i,j} | \alpha_i |^2}$$

is fulfilled, then for system (1) *the Seidel process converges to its unique solution for any choice of the initial vector.*

Proof. Suppose

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}^{(p)}=\begin{bmatrix} x_1^{(p)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(p)} \end{bmatrix}$$

are, respectively, the exact solution of system (1) and the pth approximation ($p=0, 1, 2, \ldots$) of the Seidel process for this system. We have

$$x_i=\sum_{j=1}^{i-1} \alpha_{ij}x_j+\sum_{j=i}^{n} \alpha_{ij}x_j+\beta_i$$

and

$$x_i^{(p)}=\sum_{j=1}^{i-1} \alpha_{ij}x_j^{(p)}+\sum_{j=i}^{n} \alpha_{ij}x_j^{(p-1)}+\beta_i$$

($i=1, 2, \ldots, n$), whence

$$x_i-x_i^{(p)}=\sum_{j=1}^{i-1} \alpha_{ij}(x_j-x_j^{(p)})+\sum_{j=i}^{n} \alpha_{ij}(x_j-x_j^{(p-1)})$$

and, thus,

$$|x_i-x_i^{(p)}|^2\leqslant\left\{\sum_{j=1}^{i-1} |\alpha_{ij}||x_j-x_j^{(p)}|+\sum_{j=i}^{n} |\alpha_{ij}||x_j-x_j^{(p-1)}|\right\}^2$$

Applying the Cauchy inequality (Sec. 7.7) to the sum of all terms in the braces, we obtain

$$|x_i-x_i^{(p)}|^2\leqslant s_i\left\{\sum_{j=1}^{i-1} |x_j-x_j^{(p)}|^2+\sum_{j=i}^{n} |x_j-x_j^{(p-1)}|^2\right\} \tag{2}$$

where

$$s_i=\sum_{j=1}^{n} |\alpha_{ij}|^2 \qquad (i=1, 2, \ldots, n)$$

Summing the inequalities (2) from 1 to n with respect to i, we get

$$\sum_{i=1}^{n} |x_i-x_i^{(p)}|^2\leqslant\sum_{i=1}^{n}\sum_{j=1}^{i-1} s_i|x_j-x_j^{(p)}|^2+\sum_{i=1}^{n}\sum_{j=i}^{n} s_i|x_j-x_j^{(p-1)}|^2$$

Changing the summation index in the left member and the order of summation in the right member of this inequality, we obtain

$$\sum_{j=1}^{n} |x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n-1} |x_j-x_j^{(p)}|^2\sum_{i=j+1}^{n} s_i+\sum_{j=1}^{n} |x_j-x_j^{(p-1)}|^2\sum_{i=1}^{j} s_i \tag{3}$$

Suppose

$$S_j=\sum_{i=j+1}^{n} s_i,\quad T_j=\sum_{i=1}^{j} s_i \qquad (j=1,\ 2,\ \ldots,\ n-1)$$

and

$$S_n=0,\quad T_n=\sum_{i=1}^{n} s_i$$

We evidently have

$$S_j+T_j=\sum_{i=1}^{n} s_i=\sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}|^2=\|\alpha\|_k^2<1 \qquad (j=1,\ 2,\ \ldots,\ n) \quad (4)$$

Using these notations, we can represent inequality (3) as

$$\sum_{j=1}^{n}|x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n}S_j|x_j-x_j^{(p)}|^2+\sum_{j=1}^{n}T_j|x_j-x_j^{(p-1)}|^2$$

or

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n}T_j|x_j-x_j^{(p-1)}|^2$$

On the basis of formula (4), we obtain

$$T_j=\|\alpha\|_k^2-S_j\leqslant\|\alpha\|_k^2-\|\alpha\|_k^2S_j=\|\alpha\|_k^2(1-S_j)$$

Therefore

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant\|\alpha\|_k^2\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p-1)}|^2 \quad (5)$$

From inequality (5), for $p>1$, we successively derive

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant(\|\alpha\|_k^2)^p\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(0)}|^2$$

Since $\|\alpha\|_k<1$, we then get

$$\lim_{p\to\infty}\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2=0$$

and, consequently, taking into account that $0\leqslant S_j<1$ $(j=1,2,\ldots,n)$ we obtain

$$\lim_{p\to\infty}x_j^{(p)}=x_j \qquad (j=1,\ 2,\ \ldots,\ n)$$

which is what we set out to prove.

Note. The error of iterations $x^{(p)}$ $(p=1,\ 2,\ \ldots)$ is estimated in the same way as in Sec. 9.6.

REFERENCES FOR CHAPTER 9

[1] *V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1950, Chapter II, Secs. 17 and 19 (in Russian).

Chapter 10

ESSENTIALS OF THE THEORY OF LINEAR VECTOR SPACES

10.1 THE CONCEPT OF A LINEAR VECTOR SPACE

Definition. An ordered n-tuple of numbers $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$, which, generally speaking, are complex, is called a *point* or a *vector* of n-dimensional space, while the scalars $x_1, x_2, \ldots, x_n$ are termed the *coordinates* of the vector $\boldsymbol{x}$ [1], [2], [3]. The following are vectors.

(1) The free vectors in a plane or in three-dimensional space are two-dimensional or three-dimensional vectors, respectively, in the meaning of the definition given above.

(2) Any solution of any system of linear equations in n unknowns will be an n-dimensional vector.

(3) If we have an n by m matrix (n rows and m columns), the rows are m-dimensional vectors, and the columns are n-dimensional vectors.

Two vectors $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ are considered **equal** if and only if their coordinates standing in the same positions coincide, that is, if $x_i=y_i$ for $i=1, 2, \ldots, n$.

We denote the vector $(0, 0, \ldots, 0)$ by $\mathbf{0}$ and call it the *zero vector*.

The *sum of the vectors* $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$, $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ is the vector

$$\boldsymbol{x}+\boldsymbol{y}=(x_1+y_1, x_2+y_2, \ldots, x_n+y_n)$$

whose coordinates are the sums of the respective coordinates of the vectors being added. Addition of vectors obeys the **commutative** and **associative laws**:

$$(1) \qquad \boldsymbol{x}+\boldsymbol{y}=\boldsymbol{y}+\boldsymbol{x}$$

$$(2) \qquad (\boldsymbol{x}+\boldsymbol{y})+\boldsymbol{z}=\boldsymbol{x}+(\boldsymbol{y}+\boldsymbol{z})$$

The difference between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined in similar fashion. A vector $-\boldsymbol{x}$ which satisfies the condition $(-\boldsymbol{x})+\boldsymbol{x}=\mathbf{0}$ is called the negative of the vector $\boldsymbol{x}$. It can easily be shown that

$$\boldsymbol{x}-\boldsymbol{y}=\boldsymbol{x}+(-\boldsymbol{y})$$

The *product of a vector* $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ by a scalar k is the vector

$$k\boldsymbol{x}=(kx_1, kx_2, \ldots, kx_n)$$

From this definition follow the properties of the product of a vector by a scalar:

$$\begin{aligned}
&(1) \qquad & k(\boldsymbol{x}\pm\boldsymbol{y}) &= k\boldsymbol{x}\pm k\boldsymbol{y},\\
&(2) \qquad & (k\pm l)\,\boldsymbol{x} &= k\boldsymbol{x}\pm l\boldsymbol{x},\\
&(3) \qquad & k(l\boldsymbol{x}) &= (kl)\,\boldsymbol{x},\\
&(4) \qquad & 0\boldsymbol{x} &= \boldsymbol{0},\\
&(5) \qquad & 1\boldsymbol{x} &= \boldsymbol{x},\\
&(6) \qquad & (-1)\,\boldsymbol{x} &= -\boldsymbol{x}
\end{aligned}$$

where k and l are arbitrary scalars and $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors.

For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ it is natural to define the *linear combination*

$$\alpha\boldsymbol{x}+\beta\boldsymbol{y},$$

where α, β are scalars, as a vector with coordinates $\alpha x_j+\beta y_j$ ($j=1, 2, \ldots n$).

Any collection of n-dimensional vectors which is closed under the operations of addition of vectors and the multiplication of a vector by a scalar is called a *linear vector space*. As an example, the set of all n-dimensional vectors forms an n-dimensional vector space E_n.

10.2 THE LINEAR DEPENDENCE OF VECTORS

Definition 1. The vectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ of a space E_n are termed *linearly dependent* if there exist scalars $c_1, c_2, \ldots, c_m$, not all zero, such that

$$c_1\boldsymbol{x}^{(1)}+c_2\boldsymbol{x}^{(2)}+\ldots+c_m\boldsymbol{x}^{(m)}=\boldsymbol{0} \tag{1}$$

Assuming $c_m\neq 0$, from equation (1) we have

$$\boldsymbol{x}^{(m)}=\gamma_1\boldsymbol{x}^{(1)}+\gamma_2\boldsymbol{x}^{(2)}+\ldots+\gamma_{m-1}\boldsymbol{x}^{(m-1)}$$

where

$$\gamma_j=-\frac{c_j}{c_m} \qquad (j=1, 2, \ldots, m-1)$$

Thus, *the given vectors are linearly dependent if and only if one of them is a linear combination of the others.*

But if (1) is possible in the unique case where $c_1=c_2=\ldots=c_m=0$, then the vectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ are called linearly independent; that is, the vectors are *linearly independent if and only if no linear combination of them (not all coefficients of which are zero) is* *a zero vector*. There must clearly not be any zero vector among linearly independent vectors.

Example 1. For the case of a three-dimensional vector space, E_3, the linear dependence of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ means that they are parallel to some straight line, and the linear dependence of three vectors $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, that they are parallel to some plane.

Note that if some of the vectors are linearly dependent then the whole set of vectors is also linearly dependent.

Suppose we have a collection of vectors

$$\boldsymbol{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}) \quad (j = 1, 2, \ldots, m)$$

Then, in order to determine the constants c_k $(k = 1, 2, \ldots, m)$ we get the following system by virtue of equation (1):

$$\left.\begin{array}{l} c_1x_1^{(1)} + c_2x_1^{(2)} + \ldots + c_mx_1^{(m)} = 0, \\ c_1x_2^{(1)} + c_2x_2^{(2)} + \ldots + c_mx_2^{(m)} = 0, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ c_1x_n^{(1)} + c_2x_n^{(2)} + \ldots + c_mx_n^{(m)} = 0 \end{array}\right\} \qquad (2)$$

If this system has nontrivial solutions, then the given vectors are linearly dependent, otherwise they are linearly independent.

Consider the matrix of the coordinates

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \cdot & \cdot & \cdot & \cdot \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix}$$

Let r be the rank of the matrix. Proof is given in algebra [2] that the system (2) has nontrivial solutions if and only if $r < m$. Hence, the vectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ are linearly dependent if $r < m$ and linearly independent if $r = m$ (the rank r cannot, obviously, be greater than m).

From this it follows that the rank of matrix X gives us the maximum number of linearly independent vectors contained in the given set of vectors.

Thus, if the rank of the matrix X is equal to r, then among the column vectors $\boldsymbol{x}^{(j)}$ $(j = 1, 2, \ldots, m)$: (1) there will be r linearly independent vectors, and (2) every set $r+1$ of vectors $(r+1 \leqslant m)$ of this collection are linearly dependent. The same is true of the row vectors $(x_i^{(1)}, \ldots, x_i^{(m)})$ $(i = 1, 2, \ldots, n)$ of matrix X.

Example 2. Test for linear dependence the system of vectors

$$\begin{aligned} x^{(1)} &= (1,\ -1,\ \ 1,\ -1,\ \ 1),\\ x^{(2)} &= (1,\ \ 0,\ \ 2,\ \ 0,\ \ 1),\\ x^{(3)} &= (1,\ -5,\ -1,\ \ 2,\ -1),\\ x^{(4)} &= (3,\ -6,\ \ 2,\ \ 1,\ \ 1). \end{aligned}$$

Solution. Form the matrix of the coordinates

$$X = \begin{bmatrix} 1 & 1 & 1 & 3 \\ -1 & 0 & -5 & -6 \\ 1 & 2 & -1 & 2 \\ -1 & 0 & 2 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}$$

To determine the rank r of X perform some elementary transformations: namely, subtract the sum of the first three columns from the fourth column to get

$$X \backsim \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & 0 & -5 & 0 \\ 1 & 2 & -1 & 0 \\ -1 & 0 & -2 & 0 \\ 1 & 1 & -1 & 0 \end{bmatrix}^{1)}$$

From this we conclude that all determinants of fourth order of the matrix X are zero. It is clear that there are third-order minors of X different from zero. Hence, $r=3$, and since the rank of the matrix is less than the number of vectors, the vectors $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ are linearly dependent. This is evident in the given case since

$$x^{(1)} + x^{(2)} + x^{(3)} - x^{(4)} = 0$$

Theorem 1. *The maximum number of linearly independent vectors of an n-dimensional space* E_n *is exactly equal to the dimensionality of that space.*

Proof. First of all, the space E_n has a system of n linearly independent vectors. Such, for example, is the set of n unit vectors:

$$\begin{aligned} e_1 &= (1,\ 0,\ 0,\ \ldots,\ 0),\\ e_2 &= (0,\ 1,\ 0,\ \ldots,\ 0),\\ &\ \ldots\ldots\ldots\ldots\\ e_n &= (0,\ 0,\ 0,\ \ldots,\ 1) \end{aligned}$$

---

1) The symbol $\backsim$ is used to indicate "similar matrices". See Sec. 10.13.

Thus, if

$$c_1\boldsymbol{e}_1+c_2\boldsymbol{e}_2+\ldots+c_n\boldsymbol{e}_n=(c_1,\ c_2,\ \ldots,\ c_n)=0$$

then, obviously, $c_1=c_2=\ldots=c_n=0$.

We will show that if the number of vectors $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$, $\ldots$, $\boldsymbol{x}^{(m)}$ is greater than n $(m>n)$, then they must definitely be linearly dependent. Indeed, the matrix of the coordinates of these vectors has the dimensions $n\times m$ and, consequently, its rank $r\leqslant\min(n,\ m)=n<m$, whence it follows that these vectors are linearly dependent.

Definition 2. Any set of n linearly independent vectors of an n-dimensional space is termed a *basis* of that space.

Theorem 2. *Every vector of an* n*-dimensional space* E_n *can be represented uniquely in the form of a linear combination of the vectors of a basis.*

Proof. Let $\boldsymbol{x}\in E_n$ and $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\ldots$, $\boldsymbol{\varepsilon}_n$ be a basis of E_n. By Theorem 1, the vectors $\boldsymbol{x}$, $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\ldots$, $\boldsymbol{\varepsilon}_n$ are linearly dependent, that is

$$c_0\boldsymbol{x}+c_1\boldsymbol{\varepsilon}_1+c_2\boldsymbol{\varepsilon}_2+\ldots+c_n\boldsymbol{\varepsilon}_n=\boldsymbol{0} \tag{3}$$

where a certain coefficient $c_j\neq 0$ $(0\leqslant j\leqslant n)$.

In (3) the coefficient $c_0\neq 0$, since otherwise we would have

$$c_1\boldsymbol{\varepsilon}_1+c_2\boldsymbol{\varepsilon}_2+\ldots+c_n\boldsymbol{\varepsilon}_n=\boldsymbol{0}$$

where $c_j\neq 0$ $(j\geqslant 1)$, which contradicts the linear independence of the vectors $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\ldots$, $\boldsymbol{\varepsilon}_n$. Thus, we can solve (3) for $\boldsymbol{x}$:

$$\boldsymbol{x}=\xi_1\boldsymbol{\varepsilon}_1+\xi_2\boldsymbol{\varepsilon}_2+\ldots+\xi_n\boldsymbol{\varepsilon}_n \tag{4}$$

where

$$\xi_1=-\frac{c_1}{c_0},\quad \xi_2=-\frac{c_2}{c_0},\quad \ldots,\quad \xi_n=-\frac{c_n}{c_0}$$

Thus, any vector $\boldsymbol{x}$ of the space E_n is a linear combination of the basis vectors. The expansion (4) is unique. Indeed, if there is another expansion

$$\boldsymbol{x}=\xi_1'\boldsymbol{\varepsilon}_1+\xi_2'\boldsymbol{\varepsilon}_2+\ldots+\xi_n'\boldsymbol{\varepsilon}_n \tag{4'}$$

different from the first, then, subtracting (4′) from (4), we get

$$\boldsymbol{0}=(\xi_1-\xi_1')\,\boldsymbol{\varepsilon}_1+(\xi_2-\xi_2')\,\boldsymbol{\varepsilon}_2+\ldots+(\xi_n-\xi_n')\,\boldsymbol{\varepsilon}_n \tag{5}$$

where at least one of the coefficients $\xi_j-\xi_j'\neq 0$. Equation (5) is impossible because the vectors of a basis are linearly independent. Hence, there is only one expansion of the form (4).

Geometric illustration. For the case of a three-dimensional space, formula (4) is equivalent to a decomposition of the vector $\boldsymbol{x}$ along the directions of three given vectors ε_1, ε_2 and ε_3 in the standard position (Fig. 49).

Definition 3. If $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ form a basis of n-dimensional space and

$$\boldsymbol{x} = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n$$

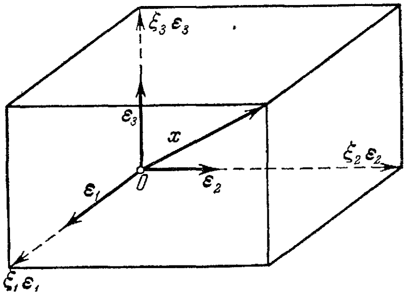


Fig. 49

then the scalars $\xi_1, \xi_2, \ldots, \xi_n$ are called the *coordinates* of the vector $\boldsymbol{x}$ in the given basis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Note that the coordinates of the vector

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

are its coordinates in the basis of unit vectors

$$\boldsymbol{e}_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj}) \qquad (j = 1, 2, \ldots, n)$$

where δ_{nj} is the Kronecker delta. We thus have the basic expansion

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \ldots + x_n\boldsymbol{e}_n \tag{6}$$

The basis of unit vectors $\boldsymbol{e}_j$ $(j = 1, 2, \ldots, n)$ will be called the *initial basis* of the space.

Definition 4. A set E_k of vectors of an n-dimensional space E_n is called a *linear subspace* of E_n if the following conditions are met:

(1) from $\boldsymbol{x} \in E_k$ and $\boldsymbol{y} \in E_k$ follows $\boldsymbol{x} + \boldsymbol{y} \in E_k$;

(2) from $\boldsymbol{x} \in E_k$ follows $\alpha\boldsymbol{x} \in E_k$ where α is any scalar. In particular, $\boldsymbol{0} \in E_k$.

Consequently, E_k may also be regarded as a vector space. The maximum number k of linearly independent vectors in E_k is called the *dimensionality* of the subspace.

From Theorem 1 it follows that $k \leqslant n$. Thus, the space E_n can have as subspaces: E_1 of one dimension, E_2 of two dimensions, and so forth up to E_n of n dimensions (the space itself). The zero vector $\boldsymbol{0}$ may be regarded as a space of zero dimensions.

Example 3. In ordinary three-dimensional space, E_3, the subspace E_1 of one dimension is a **straight line**; the subspace E_2 of two dimensions is a **plane** (Fig. 50).

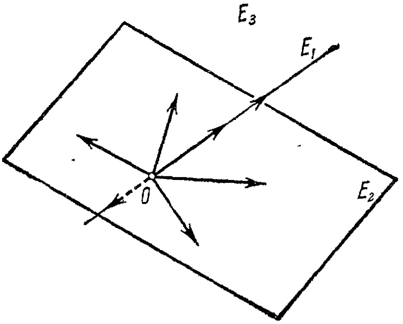


Fig. 50

Theorem 3. *If* $z_1, z_2, \ldots, z_k$ *are vectors of* E_n, *then the total set of vectors*

$$x = a_1z_1 + a_2z_2 + \ldots + a_kz_k \tag{7}$$

where a_j $(j = 1, 2, \ldots, k)$ *are arbitrary scalars, is a subspace of* E_n, *and if the vectors* $z_1, z_2, \ldots, z_k$ $(k \leqslant n)$ *are linearly independent, then the dimensionality of this subspace is equal to* k.

Conversely, any subspace E_k *of space* E_n *coincides with the set of all linear combinations of linearly independent vectors* $z_1, z_2, \ldots, z_k$ *of that subspace (basis vectors).*

Proof. The validity of the first assertion of the theorem can be verified directly.

Let us prove the second assertion. Let $x \in E_k$ and suppose x is not a linear combination of the basis vectors $z_1, z_2, \ldots, z_k$. Then, obviously, the vectors $x_1, z_1, z_2, \ldots, z_k$ are linearly independent and, thus, space E_k has $k+1$ linearly independent vectors. But this cannot be, since, by hypothesis, the maximum number of linearly independent vectors of E_k is k.

Hence, for some choice of the scalars $a_1, a_2, \ldots, a_k$ we have

$$x = a_1z_1 + a_2z_2 + \ldots + a_kz_k$$

which is what we set out to prove.

Corollary. The collection of vectors x defined by formula (7) is the smallest linear space containing the vectors $z_1, z_2, \ldots, z_k$ (it is called the *space generated by*, or *spanned by*, *the vectors* $z_1, z_2, \ldots, z_k$).

10.3 THE SCALAR PRODUCT OF VECTORS

Suppose we have the vectors

$$\boldsymbol{x}=(x_1, x_2, \ldots, x_n) \text{ and } \boldsymbol{y}=(y_1, y_2, \ldots, y_n)$$

in an n-dimensional space E_n. We assume the coordinates of the vectors to be complex numbers:

$$x_j=\xi_j+i\xi'_j, \quad y_j=\eta_j+i\eta'_j$$

where $i^2=-1$, $j=1, 2, \ldots, n$.

We introduce the conjugate quantities

$$x_j^*=\xi_j-i\xi'_j, \quad y_j^*=\eta_j-i\eta'_j$$

Then we obviously have

$$x_jx_j^*=|x_j|^2$$

By a *scalar product* of two vectors we mean a number (scalar)

$$(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n} x_jy_j^* \tag{1}$$

A scalar product has the following properties.

1. **The property of positive definiteness.** The scalar product of a vector into itself is a nonnegative scalar equal to zero if and only if the vector is zero. Indeed, from formula (1) we have

$$(\boldsymbol{x}, \boldsymbol{x})=\sum_{j=1}^{n} x_jx_j^*=\sum_{j=1}^{n}|x_j|^2\geqslant 0$$

Clearly, $(\mathbf{0}, \mathbf{0})=0$. Conversely, if $(\boldsymbol{x}, \boldsymbol{x})=0$, then $x_j=0$ ($j=1, 2, \ldots, n$) and, hence, $\boldsymbol{x}=\mathbf{0}$.

2. **Hermitian symmetry.** If two factors are interchanged, the scalar product is replaced by its conjugate. True enough, using the theorems on the conjugate quantity of a sum and of a product,[1] we have

$$(\boldsymbol{y}, \boldsymbol{x})=\sum_{j=1}^{n} y_jx_j^*=\sum_{j=1}^{n} x_j^*y_j=\left(\sum_{j=1}^{n} x_jy_j^*\right)^*=(\boldsymbol{x}, \boldsymbol{y})^*$$

Hence

$$(\boldsymbol{y}, \boldsymbol{x})=(\boldsymbol{x}, \boldsymbol{y})^* \tag{2}$$

---

[1] Here, we take advantage of the following theorems:

(a) the conjugate of a sum is equal to the sum of the conjugate quantities of the terms;

(b) the conjugate of a product is equal to the product of the conjugate quantities of the factors.

3. A **scalar factor** in the first position can be taken outside the sign of the scalar product:

$$(\alpha \boldsymbol{x},\ \boldsymbol{y}) = \alpha (\boldsymbol{x},\ \boldsymbol{y}) \tag{3}$$

The proof of this property follows directly from formula (1).

Corollary. The scalar factor occupying the second position may be taken outside the sign of the scalar product and replaced by its conjugate. We have

$$(\boldsymbol{x},\ \alpha\boldsymbol{y}) = (\alpha\boldsymbol{y},\ \boldsymbol{x})^* = [\alpha(\boldsymbol{y},\ \boldsymbol{x})]^* = \alpha^*(\boldsymbol{y},\ \boldsymbol{x})^* = \alpha^*(\boldsymbol{x},\ \boldsymbol{y})$$

and so

$$(\boldsymbol{x},\ \alpha\boldsymbol{y}) = \alpha^*(\boldsymbol{x},\ \boldsymbol{y})$$

4. **Distributivity.** If the first or second vector is the sum of two vectors, then the scalar product of this vector is the sum of the corresponding scalar products of the summands of the vector. Suppose

$$\boldsymbol{x} = \boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)}$$

where $\boldsymbol{x}^{(k)} = (x_1^{(k)},\ x_2^{(k)},\ \ldots,\ x_n^{(k)}) \quad (k=1,\ 2)$.

Proceeding from the definition of a sum of vectors, we get, by formula (1),

$$(\boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)},\ \boldsymbol{y}) = \sum_{j=1}^{n} (x_j^{(1)} + x_j^{(2)})\, y_j^* = \sum_{j=1}^{n} x_j^{(1)} y_j^* + \sum_{j=1}^{n} x_j^{(2)} y_j^* = \\ = (\boldsymbol{x}^{(1)},\ \boldsymbol{y}) + (\boldsymbol{x}^{(2)},\ \boldsymbol{y})$$

That is,

$$(\boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)},\ \boldsymbol{y}) = (\boldsymbol{x}^{(1)},\ \boldsymbol{y}) + (\boldsymbol{x}^{(2)},\ \boldsymbol{y}) \tag{4}$$

Furthermore,

$$(\boldsymbol{x},\ \boldsymbol{y}^{(1)} + \boldsymbol{y}^{(2)}) = (\boldsymbol{y}^{(1)} + \boldsymbol{y}^{(2)},\ \boldsymbol{x})^* = (\boldsymbol{y}^{(1)},\ \boldsymbol{x})^* + (\boldsymbol{y}^{(2)},\ \boldsymbol{x})^* = \\ = (\boldsymbol{x},\ \boldsymbol{y}^{(1)}) + (\boldsymbol{x},\ \boldsymbol{y}^{(2)}) \tag{5}$$

Formulas (4) and (5) may readily be extended to any finite number of vectors, namely:

$$\left(\sum_{j=1}^{m} \boldsymbol{x}^{(j)},\ \sum_{k=1}^{l} \boldsymbol{y}^{(k)}\right) = \sum_{j=1}^{m} \sum_{k=1}^{l} (\boldsymbol{x}^{(j)},\ \boldsymbol{y}^{(k)})$$

Besides the *n-dimensional complex space* that we introduced, it is useful to consider an *n-dimensional real space* consisting of a set of vectors with real coordinates.

In an n-dimensional real space, a scalar product is equal to the sum of the products of the appropriate coordinates of the vectors

$$(\boldsymbol{x},\ \boldsymbol{y}) = \sum_{j=1}^{n} x_j y_j \tag{1'}$$

The properties of a scalar product given above are then formulated as:

(1) $(x, x) \geqslant 0$ and if $(x, x)=0$, then $x=\mathbf{0}$,
(2) $(x, y)=(y, x)$,
(3) $(\alpha x, y)=(x, \alpha y)=\alpha(x, y)$ (α real),
(4) $(x+y, z)=(x, z)+(y, z)$,
$(x, y+z)=(x, y)+(x, z)$.

Using the scalar product, we can define the basic metric concepts in an n-dimensional space: length of a vector and the angle between two vectors.

1. **Length of a vector.** The length of a vector in an n-dimensional space is the nonnegative scalar

$$|x|=+\sqrt{(x, x)}$$

This definition is clearly in agreement with the notion of the length of a vector in three-dimensional space.

2. **The angle between two vectors.** The angle φ between two vectors x and y is that angle (between 0° and 180°) for which

$$\cos \varphi=\frac{(x, y)}{|x||y|}$$

For vectors in three-dimensional space, this definition is in agreement with the ordinary expression of the angle between two vectors in terms of the scalar product. We can prove that the following inequality holds true [1]:

$$|(x, y)| \leqslant |x||y|$$

For this reason, the angle between two vectors in real space is real.

10.4 ORTHOGONAL SYSTEMS OF VECTORS

Definition 1. Two vectors x and y in E_n are called *orthogonal* if their scalar product is equal to zero:

$$(x, y)=0 \tag{1}$$

If the vectors are nonzero, then orthogonality signifies that the angle between them is equal to $\frac{\pi}{2}$. The zero vector is plainly orthogonal to any vector in the space.

Thus, orthogonality is a generalized property of perpendicularity.

Definition 2. A set of vectors $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ is termed *orthogonal* if any two vectors of the set are orthogonal to each other, that is,

$$(x^{(j)}, x^{(k)}) = 0 \quad \text{for} \quad j \neq k$$

It will be noted that if the vector $x^{(1)}$ is orthogonal to the vectors $x^{(2)}, \ldots, x^{(m)}$, then this vector is orthogonal also to any linear combination of these vectors; in other words, the vector $x^{(1)}$ is orthogonal to the space spanned by the vectors $x^{(2)}, \ldots, x^{(m)}$. Indeed, if

$$(x^{(1)}, x^{(k)}) = 0 \quad \text{for} \quad k = 2, \ldots, m$$

then we have

$$\left(x^{(1)}, \sum_{k=2}^{m} c_k x^{(k)}\right) = \sum_{k=2}^{m} c_k^* (x^{(1)}, x^{(k)}) = 0$$

where $c_2, \ldots, c_m$ are arbitrary constants.

Theorem. *Nonzero pairwise orthogonal vectors* $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ *are linearly independent.*

Proof. Let

$$c_1 x^{(1)} + c_2 x^{(2)} + \ldots + c_m x^{(m)} = 0 \tag{2}$$

Form the scalar product of both members of (2) by $x^{(1)}$; we get

$$c_1^* (x^{(1)}, x^{(1)}) + c_2^* (x^{(1)}, x^{(2)}) + \ldots + c_m^* (x^{(1)}, x^{(m)}) = 0$$

or, since

$(x^{(1)}, x^{(1)}) \neq 0$ and $(x^{(1)}, x^{(j)}) = 0$ for $j \neq 1$, then $c_1^* = 0$ and $c_1 = 0$.

In exactly the same way we prove that $c_2 = 0, \ldots, c_m = 0$. Hence, the vectors $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ are linearly independent.

Corollary. An orthogonal system in n-dimensional space, E_n, has at most n vectors.

Definition 3. A basis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ of E_n is termed *orthogonal* if the basis vectors are pairwise orthogonal; that is,

$$(\varepsilon_j, \varepsilon_k) = 0 \quad \text{for} \quad j \neq k \qquad (j, k = 1, 2, \ldots, n)$$

If, moreover, the vectors ε $(j_j = 1, 2, \ldots, n)$ are unit vectors, then the orthogonal basis is called a *normalized orthogonal basis* (an *orthonormal basis*, for short). We then have

$$(\varepsilon_j, \varepsilon_k) = \delta_{jk}$$

where δ_{jk} is the Kronecker delta.

It is easy to see that the simplest orthonormal basis of E_n space is the system of unit vectors

$$\begin{array}{l} e_1 = (1,\ 0,\ 0,\ \ldots,\ 0), \\ e_2 = (0,\ 1,\ 0,\ \ldots,\ 0), \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ e_n = (0,\ 0,\ 0,\ \ldots,\ 1) \end{array}$$

that constitute the initial basis.

An orthogonal basis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ can always be normalized by dividing each of the vectors ε_j by its length. The resultant vectors

$$\tilde{\varepsilon}_j^{(0)} = \frac{\varepsilon_j}{\sqrt{(\varepsilon_j, \varepsilon_j)}} \qquad (j = 1,\ 2,\ \ldots,\ n)$$

form an orthonormal basis.

Let us express the coordinates of the vector x in the orthonormal basis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. If

$$x = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n \tag{3}$$

then, multiplying (3) scalarly on the right by ε_j, we get

$$\xi_j = (x, \varepsilon_j) \qquad (j = 1,\ 2,\ \ldots,\ n) \tag{4}$$

By analogy with vector algebra, we can say that the *coordinates of a vector in an orthonormal basis are equal to the projections of the vector on the corresponding vectors of the basis.*

Squaring (3), we get

$$(x, x) = \left(\sum_{j=1}^{n} \xi_j\varepsilon_j,\ \sum_{k=1}^{n} \xi_k\varepsilon_k \right) =$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} \xi_j\xi_k^* (\varepsilon_j, \varepsilon_k) = \sum_{j=1}^{n} \xi_j\xi_j^* = \sum_{j=1}^{n} |\xi_j|^2 \tag{5}$$

Thus, *the square of the length of a vector is equal to the sum of the squares of the moduli of its projections by the basis orthonormal vectors (an analogue of the Pythagorean theorem).* In particular, if the space E_n is real, then formula (5) may be written without the modulus sign:

$$(x, x) \sum_{j=1}^{n} (\xi_j)^2 \tag{5'}$$

10.5 TRANSFORMATIONS OF THE COORDINATES OF A VECTOR UNDER CHANGES IN THE BASIS

Suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ and $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$ are two bases of one and the same linear space E_n. Each vector of the new (second) basis $\boldsymbol{\varepsilon}_j$ has, in the old (first) basis $\boldsymbol{e}_j$, certain coordinates $s_{1j}, s_{2j}, \ldots, s_{nj}$,[1] that is

$$\boldsymbol{\varepsilon}_j = s_{1j}\boldsymbol{e}_1 + s_{2j}\boldsymbol{e}_2 + \ldots + s_{nj}\boldsymbol{e}_n \qquad (j = 1, 2, \ldots, n) \tag{1}$$

The nonsingular matrix $S = [s_{ij}]$ is called the *change-of-basis matrix* from the old basis to the new basis. (The determinant $\det S \neq 0$, otherwise the vectors $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$ would be linearly dependent.) This matrix is the transpose of the matrix that specifies the transformation of the basis. Let $\boldsymbol{x}$ be a given vector. Denote by x_i the coordinates of this vector in the old basis and by ξ_i its coordinates in the new basis. Obviously,

$$\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i = \sum_{j=1}^{n} \xi_j \boldsymbol{\varepsilon}_j$$

whence, substituting into the second sum the expression (1) for $\boldsymbol{\varepsilon}_j$, we get

$$\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i = \sum_{j=1}^{n} \xi_j \sum_{i=1}^{n} s_{ij}\boldsymbol{e}_i = \sum_{i=1}^{n} \boldsymbol{e}_i \sum_{j=1}^{n} s_{ij}\xi_j$$

Thus, by virtue of the linear independence of the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, we find

$$x_i = \sum_{j=1}^{n} s_{ij}\xi_j \qquad (i = 1, 2, \ldots, n) \tag{2}$$

If we denote

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}$$

(in other words, we consider the vector $\boldsymbol{x}$ in the new coordinates as a transformed vector referred to the old basis), then relation (2) may be rewritten in the following matrix notation:

$$\boldsymbol{x} = S\xi \tag{3}$$

which is to say, *the vector in the old coordinates (basis) is equal to the change-of-basis matrix* S (or the transpose of the matrix specifying the new basis) *multiplied by the vector in the new coordinates.*

---

[1] In designating the coordinates, we put in the first position the number of the old basis vector, and in the second, that of the new basis vector.

From formula (3) we get

$$\xi = S^{-1}x \tag{4}$$

We note an important special case similar to the transformation of rectangular coordinates. Suppose the old basis $e_1, e_2, \ldots, e_n$ and the new basis $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are real and orthonormal; that is,

$$(e_i, e_j) = \delta_{ij} \tag{5}$$

and

$$(\varepsilon_i, \varepsilon_j) = \delta_{ij} \tag{5'}$$

where δ_{ij} is the Kronecker delta.

Then formula (1) implies

$$s_{ij} = (\varepsilon_j, e_i) \qquad (i, j = 1, 2, \ldots, n) \tag{6}$$

that is, the elements of the change-of-basis matrix S are *direction cosines* and can be specified by Table 24.

TABLE 24

THE COSINES OF THE ANGLES BETWEEN THE UNIT VECTORS OF TWO BASES

Unit vectors of new system	Unit vectors of old system			
	e_1	e_2	...	e_n
ε_1	s_{11}	s_{21}	...	s_{n1}
ε_2	s_{12}	s_{22}	...	s_{n2}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
ε_n	s_{1n}	s_{2n}	...	s_{nn}

Substituting expression (1) into formula (5′) we get, by formulas (5),

$$(\varepsilon_j, \varepsilon_k) = \left(\sum_{i=1}^{n} s_{ij}e_i, \sum_{i=1}^{n} s_{ik}e_i\right) = \sum_{i=1}^{n} s_{ij}s_{ik} = \delta_{jk}$$

Thus, (1) *the sums of paired products of the corresponding direction cosines of different coordinate axes of the new orthonormal system are zero*, and (2) *the sum of the squares of the direction cosines for* *each new coordinate axis is unity.* From this,

$$S'S = E \tag{7}$$

which means that *the change-of-basis matrix from one orthonormal basis to another is orthogonal* (for details concerning orthogonal matrices see Sec. 10.6).

10.6 ORTHOGONAL MATRICES

Definition. A real matrix A is called *orthogonal* if its transpose A' coincides with the inverse A^{-1}; thus

$$A' = A^{-1} \tag{1}$$

or

$$AA' = A'A = E \tag{2}$$

An orthogonal matrix has the following properties.

1. The rows (columns) of an orthogonal matrix are orthogonal in pairs.

Indeed, if $A = [a_{ij}]$, then from (2) we have

$$\sum_{k=1}^{n} a_{ik}a_{jk} = 0 \quad \text{for} \quad i \neq j$$

and

$$\sum_{k=1}^{n} a_{ki}a_{kj} = 0 \quad \text{for} \quad i \neq j$$

2. The sum of the squares of the elements of each row (column) of an orthogonal matrix is equal to unity.

From (2), for $i = j$, we obtain

$$\sum_{k=1}^{n} a_{ik}^2 = \sum_{k=1}^{n} a_{ki}^2 = 1$$

3. The determinant of an orthogonal matrix is equal to ±1.

Thus, on the basis of (2), we have

$$\det A \det A' = \det E$$

whence, since $\det A' = \det A$ and $\det E = 1$, it follows that

$$(\det A)^2 = 1$$

and, hence, that

$$\det A = \pm 1$$

4. The transpose and the inverse of an orthogonal matrix are also orthogonal matrices. This property follows directly from formulas (1) and (2).

10.7 ORTHOGONALIZATION OF MATRICES

Suppose we have a matrix with real elements,

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

We consider the columns of A as the vectors

$$\boldsymbol{a}^{(j)}=\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \qquad (j=1,\ 2,\ \dots,\ n)$$

Thus, we can write this matrix in the form

$$A=\left[\begin{array}{|c|c|c|} \boldsymbol{a}^{(1)} & \dots & \boldsymbol{a}^{(n)} \end{array}\right]$$

Theorem 1. *Any nonsingular real matrix* A *may be represented in the form of a product of a matrix with orthogonal columns by an upper triangular matrix:*

$$A=RT$$

where R *is a matrix with orthogonal columns and* T *is an upper triangular matrix with unit diagonal.*

Proof. For the sake of simplicity, we carry out the proof for the case when the order of the matrix is $n=3$. The reasoning however will be of a general nature. Let

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Write this matrix as

$$A=[\boldsymbol{a}^{(1)},\ \boldsymbol{a}^{(2)},\ \boldsymbol{a}^{(3)}]$$

where $\boldsymbol{a}^{(j)}=\begin{bmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{bmatrix}$ are column vectors.

Since the matrix A is nonsingular, the vectors $\boldsymbol{a}^{(1)}$, $\boldsymbol{a}^{(2)}$, $\boldsymbol{a}^{(3)}$ are linearly independent.

This is so because if these vectors were linearly dependent, then in $\det A$ one of the columns would be a linear combination of the other two and, hence, $\det A=0$, which is impossible.

We seek the matrix R also in the form

$$R=[\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(2)},\ \boldsymbol{r}^{(3)}]$$

where $\boldsymbol{r}^{(j)}$ $(j=1,\ 2,\ 3)$ are the required orthogonal columns.

Set

$$\boldsymbol{r}^{(1)}=\boldsymbol{a}^{(1)} \tag{1}$$

Now decompose the vector $\boldsymbol{a}^{(2)}$ into its components $t_{12}\boldsymbol{r}^{(1)}$ and $\boldsymbol{r}^{(2)}$, of which the first is directed along the vector $\boldsymbol{r}^{(1)}$, and the second is perpendicular (orthogonal) to it (Fig. 51); thus

$$\boldsymbol{a}^{(2)}=t_{12}\boldsymbol{r}^{(1)}+\boldsymbol{r}^{(2)} \tag{2}$$

where

$$(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(2)})=0 \tag{2'}$$

Similarly, decompose vector $\boldsymbol{a}^{(3)}$ into the three components $t_{13}\boldsymbol{r}^{(1)}$, $t_{23}\boldsymbol{r}^{(2)}$ and $\boldsymbol{r}^{(3)}$, of which the first two are directed, respectively, along the vector $\boldsymbol{r}^{(1)}$ and $\boldsymbol{r}^{(2)}$, and the last is perpendicular both to the vector $\boldsymbol{r}^{(1)}$ and to the vector $\boldsymbol{r}^{(2)}$ (Fig. 51); thus

$$\boldsymbol{a}^{(3)}=t_{13}\boldsymbol{r}^{(1)}+t_{23}\boldsymbol{r}^{(2)}+\boldsymbol{r}^{(3)} \tag{3}$$

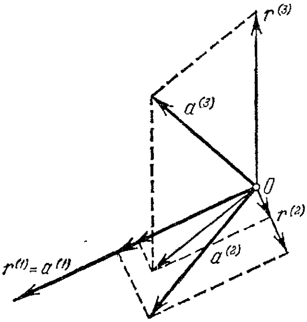


Fig. 51

where

$$(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(3)})=0 \quad \text{and} \quad (\boldsymbol{r}^{(2)},\ \boldsymbol{r}^{(3)})=0 \tag{3'}$$

From the construction it is evident that the vectors $\boldsymbol{r}^{(1)}$, $\boldsymbol{r}^{(2)}$ and $\boldsymbol{r}^{(3)}$ will be mutually perpendicular. From the systems (2) and (3), we determine the vectors $\boldsymbol{r}^{(2)}$ and $\boldsymbol{r}^{(3)}$ and also the coefficients t_{ij}. Multiplying both members of (2) scalarly by $\boldsymbol{r}^{(1)}=\boldsymbol{a}^{(1)}$, we obtain, by virtue of the orthogonality condition (2'),

$$(\boldsymbol{a}^{(2)},\ \boldsymbol{r}^{(1)})=t_{12}\,(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(1)})$$

where

$$(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(1)})\neq 0$$

Hence

$$t_{12}=\frac{(\boldsymbol{a}^{(2)},\ \boldsymbol{r}^{(1)})}{(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(1)})}$$

and

$$\boldsymbol{r}^{(2)}=\boldsymbol{a}^{(2)}-t_{12}\boldsymbol{r}^{(1)}$$

Note that since the matrix A is nonsingular, the vector $\boldsymbol{r}^{(1)}=\boldsymbol{a}^{(1)}\neq 0$ and therefore $(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(1)})\neq 0$. Moreover, $\boldsymbol{r}^{(2)}\neq 0$ since otherwise the vectors $\boldsymbol{a}^{(1)}$ and $\boldsymbol{a}^{(2)}$ would be linearly dependent.

Analogously, multiplying both members of (3) scalarly and successively by $\boldsymbol{r}^{(1)}$ and $\boldsymbol{r}^{(2)}$, we get, by virtue of the orthogonality conditions (2′) and (3′),

$$(a^{(3)}, r^{(1)}) = t_{13}(r^{(1)}, r^{(1)}),$$
$$(a^{(3)}, r^{(2)}) = t_{23}(r^{(2)}, r^{(2)})$$

From this, noting that $(r^{(1)}, r^{(1)}) \neq 0$ and $(r^{(2)}, r^{(2)}) \neq 0$, we get

$$t_{13} = \frac{(a^{(3)}, r^{(1)})}{(r^{(1)}, r^{(1)})}, \qquad t_{23} = \frac{(a^{(3)}, r^{(2)})}{(r^{(2)}, r^{(2)})}$$

and

$$r^{(3)} = a^{(3)} - t_{13}r^{(1)} - t_{23}r^{(2)}$$

It is easy to verify that the thus constructed vectors $r^{(1)}$, $r^{(2)}$ and $r^{(3)}$ are pairwise orthogonal. And so we finally have

$$\left.\begin{aligned} a^{(1)} &= r^{(1)}, \\ a^{(2)} &= t_{12}r^{(1)} + r^{(2)}, \\ a^{(3)} &= t_{13}r^{(1)} + t_{23}r^{(2)} + r^{(3)} \end{aligned}\right\} \qquad (4)$$

where

$$t_{ij} = \frac{(a^{(j)}, r^{(i)})}{(r^{(i)}, r^{(i)})} \qquad (i < j)$$

and

$$(r^{(i)}, r^{(j)}) = 0 \quad \text{for} \quad i \neq j$$

The system (4) is clearly equivalent to the matrix equation

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \cdot \begin{bmatrix} 1 & t_{12} & t_{13} \\ 0 & 1 & t_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

or

$$A = RT \qquad (5)$$

where $R = [r_{ij}]$ is a matrix with orthogonal columns and $T = [t_{ij}]$ is an upper triangular matrix with unit diagonal.

Example. Orthogonalize the columns of the matrix

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

Solution. Set

$$r^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = a^{(1)}$$

Then

$$t_{12} = \frac{(a^{(2)}, r^{(1)})}{(r^{(1)}, r^{(2)})} = \frac{1 \cdot 0 + 2 \cdot 1 + 0 \cdot 2}{0^2 + 1^2 + 2^2} = 0.4$$

We now find that

$$\mathbf{r}^{(2)}=\mathbf{a}^{(2)}-t_{12}\mathbf{r}^{(1)}=\begin{bmatrix}1\\2\\0\end{bmatrix}-0.4\begin{bmatrix}0\\1\\2\end{bmatrix}=\begin{bmatrix}1\\1.6\\-0.8\end{bmatrix}$$

To determine $\mathbf{r}^{(3)}$, compute t_{13} and t_{23} to get

$$t_{13}=\frac{(\mathbf{a}^{(3)},\ \mathbf{r}^{(1)})}{(\mathbf{r}^{(1)},\ \mathbf{r}^{(1)})}=\frac{2\cdot 0+0\cdot 1+1\cdot 2}{5}=\frac{2}{5}=0.4$$

$$t_{23}=\frac{(\mathbf{a}^{(3)},\ \mathbf{r}^{(2)})}{(\mathbf{r}^{(2)},\ \mathbf{r}^{(2)})}=\frac{2\cdot 1+0\cdot 1.6+1\cdot(-0.8)}{1^2+1.6^2+0.8^2}=\frac{1\cdot 2}{4.2}\approx 0.3$$

whence

$$\mathbf{r}^{(3)}=\mathbf{a}^{(3)}-t_{13}\mathbf{r}^{(1)}-t_{23}\mathbf{r}^{2}=\begin{bmatrix}2\\0\\1\end{bmatrix}-0.4\begin{bmatrix}0\\1\\2\end{bmatrix}-0.3\begin{bmatrix}1\\1.6\\-0.8\end{bmatrix}=\begin{bmatrix}1.70\\-0.88\\0.44\end{bmatrix}$$

Thus

$$A=\begin{bmatrix}0 & 1 & 1.7\\ 1 & 1.6 & -0.88\\ 2 & -0.8 & 0.44\end{bmatrix}\begin{bmatrix}1 & 0.4 & 0.4\\ 0 & 1 & 0.3\\ 0 & 0 & 1\end{bmatrix}$$

and the vectors

$$\mathbf{r}^{(1)}=\begin{bmatrix}0\\1\\2\end{bmatrix},\quad \mathbf{r}^{(2)}=\begin{bmatrix}1\\1.6\\-0.8\end{bmatrix},\quad \mathbf{r}^{(3)}=\begin{bmatrix}1.7\\-0.88\\0.44\end{bmatrix}$$

are pairwise orthogonal. This can be verified directly.

In certain cases it is better to orthogonalize the **rows** of a matrix, regarding them as corresponding vectors.

Suppose A' is the transpose of A and is reduced to the form

$$A'=RT \tag{6}$$

where R is a matrix with orthogonal columns and T is an upper triangular matrix with unit diagonal. Taking the transpose of (6), we get

$$A=T'R' \tag{7}$$

where T' is a lower triangular matrix and R' is a matrix with orthogonal rows. Thus, the above-described device for orthogonalizing the columns of a matrix is also suitable for orthogonalization of rows, and we have the following theorem.

Theorem 2. *Every nonsingular real matrix can be represented in the form of a product of a lower triangular matrix with unit diagonal and a matrix with orthogonal rows.*

We give yet another technique for orthogonalizing the rows of a matrix, which is sometimes of greater practical utility [5]. Suppose we have a nonsingular real matrix

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

From each ith row of A, beginning with the second, subtract the first row multiplied by a scalar λ_{i1} $(i=2, \dots, n)$ dependent on the number of the row. We then get the transformed matrix

$$A^{(1)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$$

where $a_{ij}^{(1)}=a_{ij}$ for $i=1$ and $a_{ij}^{(1)}=a_{ij}-\lambda_{i1}a_{1j}$ for $i\geqslant 2$.

Choose multipliers λ_{i1} such that the first row of matrix $A^{(1)}$ is orthogonal to all the other rows of the matrix. We have

$$\sum_{j=1}^{n} a_{1j}^{(1)}a_{ij}^{(1)}=\sum_{j=1}^{n} a_{1j}(a_{ij}-\lambda_{i1}a_{1j})=\sum_{j=1}^{n} a_{1j}a_{ij}-\lambda_{i1}\sum_{j=1}^{n} a_{1j}^{2}=0$$

whence

$$\lambda_{i1}=\frac{\sum_{j=1}^{n} a_{1j}a_{ij}}{\sum_{j=1}^{n} a_{1j}^{2}} \qquad (i=2, \dots, n)$$

Perform the same operation with matrix $A^{(1)}$, namely, leave unchanged the first two rows, and from each ith row where $i\geqslant 3$, subtract the second row of $A^{(1)}$ multiplied by a scalar λ_{i2} $(i=3, \dots, n)$. The new matrix is then

$$A^{(2)}=\begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(2)} & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}$$

where $a_{ij}^{(2)}=a_{ij}^{(1)}$ for $i=1, 2$ and $a_{ij}^{(2)}=a_{ij}^{(1)}-\lambda_{i2}a_{2j}^{(1)}$ when $i\geqslant 3$.

Since the first row of $A^{(2)}$ coincides with the first row of $A^{(1)}$ and all the other rows of $A^{(2)}$ are linear combinations of the rows of $A^{(1)}$ orthogonal to the first row of $A^{(1)}$, then the rows of $A^{(2)}$

will also be orthogonal to its first row. We choose the multipliers λ_{i2} so that the rows of $A^{(2)}$, from the third onwards, are orthogonal to the second row. This yields

$$\sum_{j=1}^{n} a_{2j}^{(2)} a_{ij}^{(2)} = \sum_{j=1}^{n} a_{2j}^{(1)} (a_{ij}^{(1)} - \lambda_{i2} a_{2j}^{(1)}) = \sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)} - \lambda_{i2} \sum_{j=1}^{n} [a_{2j}^{(1)}]^2 = 0$$

whence

$$\lambda_{i2} = \frac{\sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)}}{\sum_{j=1}^{n} [a_{2j}^{(1)}]^2} \qquad (i = 3, \ldots, n) \tag{A}$$

This process is continued till we get the matrix

$$A^{(n-1)} = \begin{bmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \ldots & a_{1n}^{(n-1)} \\ a_{21}^{(n-1)} & a_{22}^{(n-1)} & \ldots & a_{2n}^{(n-1)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(n-1)} & a_{n2}^{(n-1)} & \ldots & a_{nn}^{(n-1)} \end{bmatrix}$$

all the rows of which are orthogonal in pairs:

$$\sum_{j=1}^{n} a_{kj}^{(n-1)} a_{ij}^{(n-1)} = 0 \quad \text{when} \quad k \neq i$$

The matrix $A^{(n-1)} = \tilde{R}$ with orthogonal rows was obtained from the given matrix A as a result of a chain of elementary transformations. We therefore have the valid equality

$$\tilde{R} = \Lambda A \tag{8}$$

where Λ is a nonsingular matrix, which in our case is a lower triangular matrix.

It is easy to restore matrix Λ by taking the unit matrix E and performing all the elementary transformations carried out with respect to A. From formula (8) we finally get

$$A = \tilde{T}\tilde{R}$$

where $\tilde{T} = \Lambda^{-1}$ is a lower triangular matrix.

We give some properties of matrices with orthogonal rows.

Lemma. *If the columns of a real matrix constitute an orthogonal system of vectors, then the product of the transpose of a matrix by that matrix is equal to the diagonal matrix.*

Proof. Let $A=[a_{ij}]$ be a given matrix. It is required to prove that $A'A=D$, where $A'=[a_{ji}]$ is the transpose of matrix A and

$$D=\begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix} \quad \text{is a diagonal matrix.}$$

Assuming $D=[d_{ij}]$ we have, by the rule of matrix multiplication,

$$d_{ij}=\sum_{k=1}^{n} a_{ki}a_{kj}$$

whence, since a_{ki} are the coordinates of the ith vector $a^{(i)}$ and a_{kj} are the coordinates of the jth vector $a^{(j)}$, we get

$$d_{ij}=\sum_{k=1}^{n} a_{ki}a_{kj}=(a^{(i)},\ a^{(j)})=0 \quad \text{if} \quad i\neq j$$

Consequently, $D=[d_{ij}]$ is a diagonal matrix.

Corollary. The product of a real matrix with orthogonal rows by the transpose of that matrix is equal to the diagonal matrix, that is, $AA'=D$.

Theorem 3. *Any nonsingular real matrix A with orthogonal columns is an orthogonal matrix postmultiplied by a diagonal matrix.*

Proof. By the lemma we have

$$A'A=D \tag{9}$$

where $D=[d_{ij}]$ is a diagonal matrix. If $A=[a_{ij}]$, then obviously

$$d_{ii}=\sum_{k=1}^{n} a_{ki}^2>0$$

Put

$$\rho_i=\sqrt{d_{ii}}>0 \qquad (i=1,\ 2,\ \dots,\ n)$$

and

$$d=\begin{bmatrix} \rho_1 & 0 & \dots & 0 \\ 0 & \rho_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \rho_n \end{bmatrix}$$

It is plain that $D=d^2$. From (9) we have $A'A=d^2$, whence

$$d^{-1}A'Ad^{-1}=E$$

Since $(d^{-1})'=d^{-1}$, it follows that $(Ad^{-1})'(Ad^{-1})=E$. Consequently, matrix $Ad^{-1} = U$ is orthogonal and, hence,

$$A = Ud \tag{10}$$

which completes the proof.

Corollary. A nonsingular real matrix with orthogonal rows may be represented in the form of a product of a diagonal matrix by an orthogonal matrix.

Indeed, let A be a matrix with orthogonal rows; then A' is a matrix with orthogonal columns. By formula (10) we have $A' = Ud$, where U is an orthogonal matrix and d is a diagonal matrix which may be determined from the relation

$$AA' = d^2$$

From this we get

$$A = (A')' = d'U' = dU'$$

where U' is also an orthogonal matrix.

Note. In order to transform a given nonsingular real matrix A with orthogonal columns (rows) into an orthogonal matrix, it is sufficient to normalize the columns (rows), which means that the element of every column (row) is to be divided by the square root of the sum of the squares of the elements of that column (row). For instance, if $A = [a_{ij}]$ is a matrix with orthogonal columns, then the matrix

$$\tilde{A} = [\tilde{a}_{ij}],$$

where $\tilde{a}_{ij} = \dfrac{a_{ij}}{\sqrt{\sum\limits_{k=1}^{n} a_{kj}^2}}$ $(i, j = 1, 2, \ldots, n)$ is an orthogonal matrix.

10.8 APPLYING ORTHOGONALIZATION METHODS TO THE SOLUTION OF SYSTEMS OF LINEAR EQUATIONS

A. FIRST METHOD (ORTHOGONALIZATION OF COLUMNS)

Suppose we have a system of linear equations

$$A\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

with a nonsingular real matrix A. Orthogonalizing the columns of A we obtain a matrix R; here, $A = RT$, where T is an upper triangular matrix. We have

$$RT\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

Premultiplying both members of (2) by R', we get

$$R'RT\boldsymbol{x} = R'\boldsymbol{b} \tag{3}$$

But, as we know, $R'R = D$, where D is a diagonal matrix. Introducing the notation $R'\boldsymbol{b} = \beta$, we have

$$DT\boldsymbol{x} = \beta$$

whence

$$\boldsymbol{x} = (DT)^{-1}\beta = T^{-1}D^{-1}\beta \tag{4}$$

The matrix D^{-1}, which is the inverse of the diagonal matrix, is found without difficulty; namely, if

$$D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & d_{nn} \end{bmatrix}$$

then

$$D^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 & \dots & 0 \\ 0 & d_{22}^{-1} & \dots & 0 \\ 0 & 0 & \dots & d_{nn}^{-1} \end{bmatrix}$$

It is also relatively simple to find the inverse T^{-1} of the triangular matrix T.

Example 1. Using the method of orthogonalization of columns, solve the following system of equations:

$$\left.\begin{aligned} 0.4x_1 + 0.3x_2 - 0.2x_3 &= 2, \\ 0.6x_1 - 0.5x_2 + 0.3x_3 &= 2.5, \\ 0.3x_1 + 0.2x_2 + 0.5x_3 &= 11 \end{aligned}\right\}$$

Solution. Represent the matrix A of this system in the form of a product of matrix R with orthogonal columns by a triangular matrix with unit diagonal:

$$A = RT = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} 1 & \lambda_{12} & \lambda_{13} \\ 0 & 1 & \lambda_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Set

$$\boldsymbol{r}^{(1)} = \boldsymbol{a}^{(1)}, \quad \boldsymbol{r}^{(2)} = \boldsymbol{a}^{(2)} - \lambda_{12}\boldsymbol{r}^{(1)}, \quad \boldsymbol{r}^{(3)} = \boldsymbol{a}^{(3)} - \lambda_{13}\boldsymbol{r}^{(1)} - \lambda_{23}\boldsymbol{r}^{(2)}$$

We have

$$\boldsymbol{r}^{(1)} = \begin{bmatrix} 0.4 \\ 0.6 \\ 0.3 \end{bmatrix}$$

Using formulas (4) of the preceding section, we get

$$\lambda_{12}=\frac{(a^{(2)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})}=\frac{0.12-0.3+0.06}{0.16+0.36+0.09}=-\frac{0.12}{0.61}=-0.1967,$$

$$r^{(2)}=\begin{bmatrix}0.3\\-0.5\\0.2\end{bmatrix}+0.1967\begin{bmatrix}0.4\\0.6\\0.3\end{bmatrix}=\begin{bmatrix}0.3787\\-0.3820\\0.2590\end{bmatrix}$$

Check:

$$(r^{(1)},\ r^{(2)})=\begin{bmatrix}0.4\\0.6\\0.3\end{bmatrix}'\begin{bmatrix}0.3787\\-0.3820\\0.2590\end{bmatrix}=\left.\begin{matrix}0.1515\\-0.2292\\0.0777\end{matrix}\right\}=0,$$

$$\lambda_{13}=\frac{(a^{(3)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})}=\frac{-0.08+0.18+0.15}{0.61}=\frac{0.25}{0.61}=0.4098,$$

$$\lambda_{23}=\frac{(a^{(3)},\ r^{(2)})}{(r^{(2)},\ r^{(2)})}=-\frac{0.07574-0.11460+0.12950}{0.35}=-0.1714,$$

$$r^{(3)}=\begin{bmatrix}-0.2\\0.3\\0.5\end{bmatrix}-0.4098\begin{bmatrix}0.4\\0.6\\0.3\end{bmatrix}+0.1714\begin{bmatrix}0.3787\\-0.3820\\0.2590\end{bmatrix}=\begin{bmatrix}-0.2990\\-0.0114\\0.4215\end{bmatrix}$$

Check:

$$(r^{(1)},\quad r^{(3)})=(r^{(2)},\quad r^{(3)})=0$$

Thus,

$$A=\begin{bmatrix}\underbrace{\begin{matrix}0.4 & 0.3787 & -0.2990\\0.6 & -0.3820 & -0.0114\\0.3 & 0.2590 & 0.4215\end{matrix}}_{R}\end{bmatrix}\begin{bmatrix}\underbrace{\begin{matrix}1 & -0.1967 & 0.4098\\0 & 1 & -0.1714\\0 & 0 & 1\end{matrix}}_{T}\end{bmatrix}$$

From formula (4) we have

$$x=T^{-1}D^{-1}R'b$$

where $D=R'R$ is a diagonal matrix and

$$b=\begin{bmatrix}2\\2.5\\11\end{bmatrix}$$

For the matrix D and its inverse D^{-1} we obtain the values

$$D=\begin{bmatrix}0.61 & 0 & 0\\0 & 0.35 & 0\\0 & 0 & 0.2672\end{bmatrix}\quad\text{and}\quad D^{-1}=\begin{bmatrix}1.64 & 0 & 0\\0 & 2.81 & 0\\0 & 0 & 3.75\end{bmatrix}$$

Furthermore,

$$R'\boldsymbol{b}=\begin{bmatrix}0.4 & 0.6 & 0.3\\ 0.3787 & -0.3820 & 0.2590\\ 0.2990 & -0.0114 & 0.4215\end{bmatrix}\begin{bmatrix}2\\ 2.5\\ 11\end{bmatrix}=\begin{bmatrix}5.6\\ 2.67\\ 4.08\end{bmatrix}$$

Finally, we compute (in the usual way)

$$T^{-1}=\begin{bmatrix}1 & 0.1967 & -0.3761\\ 0 & 1 & 0.1714\\ 0 & 0 & 1\end{bmatrix}$$

The result is thus

$$\boldsymbol{x}=\begin{bmatrix}1 & 0.1967 & -0.3761\\ 0 & 1 & 0.1714\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1.64 & 0 & 0\\ 0 & 2.81 & 0\\ 0 & 0 & 3.75\end{bmatrix}\begin{bmatrix}5.6\\ 2.67\\ 4.08\end{bmatrix}=\begin{bmatrix}5.0238\\ 10.0475\\ 15.0087\end{bmatrix}$$

and, hence,

$$x_1=5.0238, \qquad x_2=10.0475, \qquad x_3=15.0087$$

The exact values of the roots are $x_1=5$, $x_2=10$, $x_3=15$.

B. SECOND METHOD (ORTHOGONALIZATION OF ROWS)

Suppose we have a system

$$A\boldsymbol{x}=\boldsymbol{b} \tag{5}$$

where $\det A \neq 0$.

Transform the rows of (5) by the device given in the preceding section so that the matrix A is transformed to the matrix R with orthogonal rows. Then the vector $\boldsymbol{b}$ will pass into some vector $\boldsymbol{\beta}$. As a result, we get the equivalent system

$$R\boldsymbol{x}=\boldsymbol{\beta} \tag{6}$$

Hence

$$\boldsymbol{x}=R^{-1}\boldsymbol{\beta} \tag{7}$$

As we know, $RR'=D=d^2$, where d is a diagonal matrix and $R=dU$, where U is an orthogonal matrix, and so

$$R^{-1}=(dU)^{-1}=U^{-1}d^{-1}=U'd'd^{-2}=(dU)'\,d^{-2}=R'd^{-2}=R'D^{-1}$$

Thus, on the basis of formula (7) we finally have

$$\boldsymbol{x}=R'D^{-1}\boldsymbol{\beta} \tag{8}$$

where

$$D=RR' \tag{9}$$

By using (8) we can avoid the most laborious process of finding the inverse of a nondiagonal matrix. The presence of D^{-1} does not complicate matters because D is a diagonal matrix. Formula (9), which is needed anyway, may also be used for a check.

Example 2. Use the method of orthogonalization of rows to solve the system of equations

$$\left.\begin{aligned} 3.00x_1+0.15x_2-0.09x_3&=6.00,\\ 0.08x_1+4.00x_2-0.16x_3&=12.00,\\ 0.05x_1+0.30x_2+5.00x_3&=20.00 \end{aligned}\right\} \qquad \text{(I)}$$

Solution. Using the formulas of the preceding section, we determine the multipliers:

$$\lambda_{21}=\frac{3.00\cdot 0.08+0.15\cdot 4.00+(-0.09)\cdot(-0.16)}{3.00^2+0.15^2+0.09^2}=\frac{0.8544}{9.0306}=0.0946,$$

$$\lambda_{31}=\frac{3.00\cdot 0.05+0.15\cdot 0.30-0.09\cdot 5.00}{3.00^2+0.15^2+0.09^2}=-\frac{0.2550}{9.0306}=-0.0282$$

Retaining the first equation of system (I), subtract from each subsequent equation the first multiplied by the corresponding multipliers λ_{i1} $(i=2,\ 3)$:

$$\left.\begin{aligned} 3.00x_1+0.15x_2-0.09x_3&=6.00\\ -0.2038x_1+3.9858x_2-0.1685x_3&=11.4324,\\ 0.1346x_1+0.3042x_2+4.9975x_3&=20.1692 \end{aligned}\right\} \qquad \text{(II)}$$

For system (II) we determine the multiplier

$$\lambda_{32}=\frac{-0.2038\cdot 0.1346+3.9858\cdot 0.3042-0.1685\cdot 4.9975}{0.2038^2+3.9858^2+0.1685^2}=\frac{0.3430}{15.9565}=0.0215$$

Retaining the first two equations of system (II), subtract from the third equation the second multiplied by λ_{32}:

$$\left.\begin{aligned} 3.00x_1+0.15x_2-0.09x_3&=6.00,\\ -0.2038x_1+3.9858x_2-0.1685x_3&=11.4324,\\ 0.1390x_1+0.2185x_2+5.0011x_3&=19.9234 \end{aligned}\right\} \qquad \text{(III)}$$

The matrix

$$R=\begin{bmatrix} 3.00 & 0.15 & -0.09\\ -0.2038 & 3.9858 & -0.1685\\ 0.1390 & 0.2185 & 5.0011 \end{bmatrix}$$

has orthogonal rows. As a check, form the matrix

$$D=RR'=\begin{bmatrix} 9.0306 & 0.0017 & -0.0002\\ 0.0017 & 15.9565 & -0.0018\\ -0.0002 & -0.0018 & 25.0780 \end{bmatrix}\approx$$

$$\approx\begin{bmatrix} 9.0306 & 0 & 0\\ 0 & 15.9565 & 0\\ 0 & 0 & 25.0780 \end{bmatrix}$$

Using formula (8), we get

$$x = R'D^{-1}\beta = \begin{bmatrix} 3.00 & -0.2038 & 0.1390 \\ 0.15 & 3.9858 & 0.2185 \\ -0.09 & -0.1685 & 5.0011 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0.1107 & 0 & 0 \\ 0 & 0.0626 & 0 \\ 0 & 0 & 0.0399 \end{bmatrix} \begin{bmatrix} 6.00 \\ 11.4324 \\ 19.9234 \end{bmatrix} = \begin{bmatrix} 1.957 \\ 3.126 \\ 3.803 \end{bmatrix}$$

Hence

$$x_1 = 1.957, \quad x_2 = 3.126, \quad x_3 = 3.803.$$

C. THIRD METHOD (METHOD OF ORTHOGONAL MATRICES)

Suppose we have a linear system reduced to the form

$$Rx = \beta \tag{10}$$

where $R = [r_{ij}]$ is a nonsingular matrix with orthogonal rows and

$$\beta = \begin{bmatrix} \beta_i \\ \vdots \\ \beta_n \end{bmatrix}$$

is the vector of constant terms.

Multiplying each equation of the system (10) by the normalizing factor

$$\mu_i = \frac{1}{\sqrt{\sum_{=1}^{n} r_{ij}^2}} \qquad (i = 1, 2, \ldots, n)$$

we get the system

$$\tilde{R}x = \tilde{\beta} \tag{11}$$

where $\tilde{R} = [\mu_i r_{ij}]$ is an orthogonal matrix and $\tilde{\beta} = \begin{bmatrix} \mu_1\beta_1 \\ \mu_2\beta_2 \\ \cdots\cdot \\ \mu_n\beta_n \end{bmatrix}$ is the new vector of constant terms.

From equation (11) we have

$$x = \tilde{R}^{-1}\tilde{\beta} = \tilde{R}'\tilde{\beta} \tag{12}$$

10.9 THE SOLUTION SPACE OF A HOMOGENEOUS SYSTEM

We now consider a homogeneous linear system of equations

$$\left.\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=0,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=0 \end{aligned}\right\} \qquad (1)$$

or, briefly,

$$A\boldsymbol{x}=\mathbf{0} \qquad (1')$$

where $A=[a_{ij}]$ and $\boldsymbol{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n \end{bmatrix}$ is the desired vector.

If $\det A\neq 0$, then, by the Cramer rule, the system (1) has a unique trivial solution $\boldsymbol{x}=\mathbf{0}$.

Let $\det A=0$. Then (1) has an infinite number of solutions (including nontrivial solutions).

From formula (1′) it follows that: (1) if $\boldsymbol{x}$ is a solution of (1′), then $c\boldsymbol{x}$, where c is an arbitrary constant, is also a solution of this equation; (2) if $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are solutions of (1′), then the sum $\boldsymbol{x}^{(1)}+\boldsymbol{x}^{(2)}$ is also a solution of (1′).

Indeed, if

$$A\boldsymbol{x}=\mathbf{0}$$

then

$$A(c\boldsymbol{x})=cA\boldsymbol{x}=\mathbf{0}$$

Hence, $c\boldsymbol{x}$ is a solution of the homogeneous system (1).

Similarly, if

$$A\boldsymbol{x}^{(1)}=\mathbf{0} \quad \text{and} \quad A\boldsymbol{x}^{(2)}=\mathbf{0}$$

then

$$A(\boldsymbol{x}^{(1)}+\boldsymbol{x}^{(2)})=A\boldsymbol{x}^{(1)}+A\boldsymbol{x}^{(2)}=\mathbf{0}+\mathbf{0}=\mathbf{0}$$

That is, $\boldsymbol{x}^{(1)}+\boldsymbol{x}^{(2)}$ is also a solution of the homogeneous system (1).

In this way we find that *any linear combination of solutions of the homogeneous system* (1) *is also a solution of the system*. Hence, the set of all solutions of (1) form a vector space which we call the *solution space*. The following theorem holds true.

Theorem. *If* n *is the number of unknowns of a homogeneous system* (1) *and* r *is the rank of its matrix* A, *then the dimensionality of the solution space* is k = n − r.

The basis of a solution space is termed the *fundamental system of solutions*. If for system (1) the fundamental system of solutions

$$\begin{aligned}
\boldsymbol{x}^{(1)} &= (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}),\\
\boldsymbol{x}^{(2)} &= (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}),\\
&\ldots\ldots\ldots\ldots\\
\boldsymbol{x}^{(k)} &= (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}),
\end{aligned}$$

is known, then all its solutions are contained in the formula

$$\boldsymbol{x} = c_1\boldsymbol{x}^{(1)} + c_2\boldsymbol{x}^{(2)} + \ldots + c_k\boldsymbol{x}^{(k)} \tag{2}$$

or, written out,

$$\left.\begin{aligned}
x_1 &= c_1x_1^{(1)} + c_2x_1^{(2)} + \ldots + c_kx_1^{(k)},\\
x_2 &= c_1x_2^{(1)} + c_2x_2^{(2)} + \ldots + c_kx_2^{(k)},\\
&\ldots\ldots\ldots\ldots\ldots\\
x_n &= c_1x_n^{(1)} + c_2x_n^{(2)} + \ldots + c_kx_n^{(k)}
\end{aligned}\right\}$$

where $c_1, c_2, \ldots, c_k$ are arbitrary constants.

To find the fundamental system of solutions, one isolates the nonzero rth order minor δ_r in the matrix A. Suppose

$$\delta_r = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1r}\\ a_{21} & a_{22} & \ldots & a_{2r}\\ \cdot & \cdot & \cdot & \cdot\\ a_{r1} & a_{r2} & \ldots & a_{rr}\end{vmatrix} \neq 0$$

This can always be done by interchanging the equations of system (1) and altering the numbering of the unknowns. It is then easy to prove that the equations of (1), beginning with the $(r+1)$th, are consequences of the first r equations of the system; that is to say, they are satisfied if the first r equations of system (1) are satisfied. It is therefore sufficient to consider the subsystem

$$\left.\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r &= -a_{1,r+1}x_{r+1} - \ldots - a_{1n}x_n,\\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2r}x_r &= -a_{2,r+1}x_{r+1} - \ldots - a_{2n}x_n,\\
&\ldots\ldots\ldots\ldots\ldots\\
a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rr}x_r &= -a_{r,r+1}x_{r+1} - \ldots - a_{rn}x_n
\end{aligned}\right\} \tag{3}$$

whose determinant δ_r is nonzero.

In the system (3) the values of the unknowns

$$x_{r+1} = c_1, \quad x_{r+2} = c_2, \quad \ldots, \quad x_n = c_{n-r} = c_k$$

may be considered arbitrary. Solving (3) for the unknowns x_1, x_2,

$\ldots, x_r$, we get

$$\left.\begin{aligned} x_1 &= \alpha_{11}c_1 + \alpha_{12}c_2 + \ldots + \alpha_{1k}c_k, \\ x_2 &= \alpha_{21}c_1 + \alpha_{22}c_2 + \ldots + \alpha_{2k}c_k, \\ &\cdots\cdots\cdots\cdots \\ x_r &= \alpha_{r1}c_1 + \alpha_{r2}c_2 + \ldots + \alpha_{rk}c_k \end{aligned}\right\} \qquad (4)$$

where α_{ij} $(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, k)$ are quite definite constants. Moreover,

$$\left.\begin{aligned} x_{r+1} &= c_1, \\ x_{r+2} &= c_2, \\ &\cdots \\ x_n &= c_k \end{aligned}\right\} \qquad (4')$$

Formulas (4) and (4′) yield the complete system of solutions of system (1). For the fundamental system of solutions we can take

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} \alpha_{11} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{r1} \\ 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{bmatrix} \alpha_{12} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{r2} \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \ldots, \quad \boldsymbol{x}^{(k)} = \begin{bmatrix} \alpha_{1k} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{rk} \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}.$$

These solutions can be found directly from the system (3) if in (3) we successively set

$$\begin{gathered} x_{r+1} = 1,\ x_{r+2} = \ldots = x_n = 0, \\ x_{r+1} = 0,\ x_{r+2} = 1,\ x_{r+3} = \ldots = x_n = 0, \\ \cdots\cdots\cdots\cdots\cdots, \\ x_{r+1} = \ldots = x_{n-1} = 0,\ x_n = 1 \end{gathered}$$

Example. Find the fundamental system of solutions of the following system of homogeneous equations:

$$\left.\begin{aligned} x_1 - x_2 + 5x_3 - x_4 &= 0, \\ x_1 + x_2 - 2x_3 + 3x_4 &= 0, \\ 3x_1 - x_2 + 8x_3 + x_4 &= 0, \\ x_1 + 3x_2 - 9x_3 + 7x_4 &= 0 \end{aligned}\right\} \qquad (5)$$

Solution. The rank of the matrix of system (5) $r=2$, and

$$\delta=\begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix}=2\neq 0$$

For this reason, the last two equations of (5) are consequences of the first two. We solve the subsystem

$$\left.\begin{aligned} x_1-x_2 &= -5x_3+x_4, \\ x_1+x_2 &= 2x_3-3x_4 \end{aligned}\right\}$$

First assuming $x_3=1$, $x_4=0$, and then $x_3=0$ and $x_4=1$, we get two linearly independent solutions

$$\boldsymbol{x}^{(1)}=\left(-\frac{3}{2},\ \frac{7}{2},\ 1,\ 0\right),$$

$$\boldsymbol{x}^{(2)}=(-1,\ -2,\ 0,\ 1)$$

which form the fundamental system of solutions of the system (5).

The vectors $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ form a basis of the solution space of the given system, and all its solutions are determined by the formulas

$$\left.\begin{aligned} x_1 &= -3c_1-c_2, \\ x_2 &= 7c_1-2c_2, \\ x_3 &= 2c_1, \\ x_4 &= \phantom{-3c_1-{}} c_2 \end{aligned}\right\}$$

where c_1 and c_2 are arbitrary constants (for the sake of convenience, the first constant is taken in the form $2c_1$).

10.10 LINEAR TRANSFORMATIONS OF VARIABLES

Suppose $x_1, x_2, \ldots, x_n$ are a set of variables and $y_1, y_2, \ldots, y_n$ are another set of variables connected with the first set by the relations

$$\left.\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n), \\ y_2 &= f_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ y_n &= f_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \quad (1)$$

where $f_1, f_2, \ldots, f_n$ are given functions.

We will use the term *transformation* for a transition from the set $x_1, x_2, \ldots, x_n$ to the set $y_1, y_2, \ldots, y_n$.

Definition. The transformation (1) is called *linear* if the new variables $y_1, y_2, \ldots, y_n$ are homogeneous linear functions of the old variables $x_1, x_2, \ldots, x_n$; that is,

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{aligned}\right\} \qquad (2)$$

where a_{ij} ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n$) are constants.

The linear transformation (2) is uniquely defined by its square matrix of coefficients (*transformation matrix*)

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$

Conversely, having a square matrix $A = [a_{ij}]$, one can construct a linear transformation for which this matrix will serve as the matrix of coefficients. There thus exists a *one-to-one correspondence* between linear transformations and square matrices.

The sets of variables $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ may be regarded as column vectors

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix},$$

of the same vector space E_n. For this reason, the transformation (2) maps the space E_n onto itself or onto a proper subset of itself.

Example 1. The transformation

$$\left.\begin{aligned} y_1 &= x_1 + x_2, \\ y_2 &= x_1 + x_2 \end{aligned}\right\}$$

carries the space E_2 into the subspace $y_1 = y_2$ of dimensionality 1.

The relations (2) are equivalent to a single matrix relation,

$$\boldsymbol{y} = A\boldsymbol{x} \qquad (3)$$

Indeed, by the rule for multiplying matrices we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{bmatrix}$$

From this, via the notion of the equality of matrices, we get formulas (2). By formula (3), matrix A may be regarded as a linear-transformation *operator*.

It is easy to see that a linear transformation has two basic properties:

(1) a constant factor can be taken outside the sign of the linear-transformation operator:

$$A(\alpha \boldsymbol{x}) = \alpha A \boldsymbol{x}$$

(2) the linear-transformation operator of a sum of several vectors is equal to the sum of the operators of these vectors; thus

$$A(\boldsymbol{x}+\boldsymbol{z}) = A\boldsymbol{x} + A\boldsymbol{z}$$

As a consequence, we have

$$A(\alpha\boldsymbol{x}+\beta\boldsymbol{z}) = \alpha A\boldsymbol{x} + \beta A\boldsymbol{z}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are vectors, α and β are scalars.

Example 2. We have a plane Ox_1x_2. Suppose each vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is associated with a vector

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

which is the projection of vector $\boldsymbol{x}$ on the x_1-axis (*projection transformation*) (Fig. 52). Show that the given transformation is a linear transformation and find the transformation matrix.

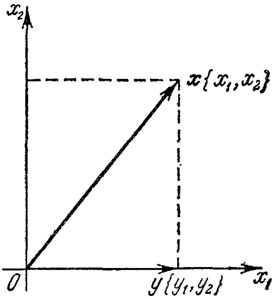


Fig. 52

Solution. We obviously have

$$\left.\begin{aligned} y_1 &= x_1, \\ y_2 &= 0 \end{aligned}\right\}$$

and consequently the projection transformation is linear. The transformation matrix is

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Let us investigate the meaning of the elements a_{ij} of A. We consider the unit vectors directed along the coordinate axes Ox_1,

$Ox_2, \ldots, Ox_n$:

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \ldots, \quad e_n = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

Applying the transformation A to e_j, we have

$$Ae_j = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \qquad (j = 1, 2, \ldots, n)$$

Thus, a_{ij} is the ith coordinate of the transformed jth unit vector.

Example 3. Suppose every radius vector $\boldsymbol{x}$ in the x_1x_2-plane is replaced by a radius vector $\boldsymbol{y}$ of the same length turned through an angle α from $\boldsymbol{x}$ (*rotation transformation*) (Fig. 53).

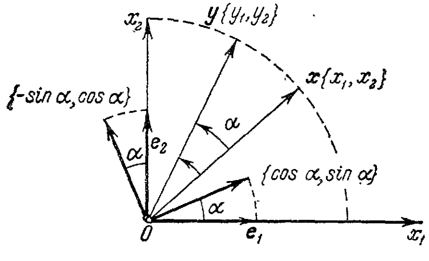


Fig. 53

Show that the given transformation is linear and find the transformation matrix.

Solution. Connect vector $\boldsymbol{y}$ with a coordinate system Oy_1y_2, which is turned through an angle α relative to the coordinate system Ox_1x_2. Since the coordinates of the vector $\boldsymbol{y}$ in the system Oy_1y_2 are clearly x_1 and x_2, the coordinates of this vector in the old Ox_1x_2 system are given, by familiar formulas of analytic geometry, by

$$\left.\begin{aligned} y_1 &= x_1 \cos\alpha - x_2 \sin\alpha, \\ y_2 &= x_1 \sin\alpha + x_2 \cos\alpha \end{aligned}\right\} \qquad (4)$$

The rotation transformation is thus linear and its matrix is of the form

$$A=\begin{bmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{bmatrix}$$

The problem can be solved differently: as a result of the transformation, the unit vector $\boldsymbol{e}_1$ will plainly go into the vector $\{\cos\alpha,\ \sin\alpha\}$ and the unit vector $\boldsymbol{e}_2$ will go into the vector $\{-\sin\alpha,\ \cos\alpha\}$. Hence, by the foregoing, the transformation matrix is

$$A=\begin{bmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{bmatrix}$$

and we again arrive at formulas (4).

We now define the following operations with the linear transformations A and B: (1) **addition** $A+B$, (2) **multiplication by a scalar** αA, (3) **composition** BA:

(1) $$(A+B)\,\boldsymbol{x}=A\boldsymbol{x}+B\boldsymbol{x} \tag{5}$$

(2) $$(\alpha A)\,\boldsymbol{x}=\alpha\,(A\boldsymbol{x}) \tag{5$'$}$$

(3) $$(BA)\,\boldsymbol{x}=B\,(A\boldsymbol{x}) \tag{5$''$}$$

where $\boldsymbol{x}$ is a vector.

It can readily be shown that if A and B are linear transformations, then the resulting transformations $A+B$, αA, and BA are also linear.

The following principle holds true: *each operation from* (1) *to* (3) *involving linear transformations is associated with the same operation on the matrices of these linear transformations*; this is taken to mean that the formulas (5), (5′) and (5″) are to be regarded in their literal meaning.

For the proof, we confine ourselves to the third case (5″) [formulas (5) and (5′) are verified more simply].

Suppose it is required to perform two linear transformations in succession:

$$\left.\begin{aligned} y_1&=a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n,\\ y_2&=a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ y_n&=a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n \end{aligned}\right\} \tag{6}$$

with matrix $A=[a_{ij}]$ and

$$\left.\begin{aligned} z_1&=b_{11}y_1+b_{12}y_2+\ldots+b_{1n}y_n,\\ z_2&=b_{21}y_1+b_{22}y_2+\ldots+b_{2n}y_n,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ z_n&=b_{n1}y_1+b_{n2}y_2+\ldots+x_{nn}y_n \end{aligned}\right\} \tag{7}$$

with matrix $B=[b_{ij}]$.

Denote by $C=[c_{i}]$ the matrix of the composition of these transformations in the indicated order, that is, going from the variables $x_1, x_2, \dots, x_n$ to the variables $z_1, z_2, \dots, z_n$. Writing the transformations (6) and (7) compactly as

$$y_k=\sum_{j=1}^{n} a_{kj}x_j \qquad (k=1, 2, \dots, n), \tag{6'}$$

$$z_i=\sum_{k=1}^{n} b_{ik}y_k \qquad (i=1, 2, \dots, n) \tag{7'}$$

and substituting formula (6′) into (7′), we get

$$z_i=\sum_{k=1}^{n} b_{ik}\left(\sum_{j=1}^{n} a_{kj}x_j\right)=\sum_{i=1}^{n} x_j\sum_{k=1}^{n} b_{ik}a_{kj} \tag{8}$$

Thus, the coefficient of x_j in the expression for z_i, or element c_{ij} of the matrix C, is of the form

$$c_{ij}=\sum_{k=1}^{n} b_{ik}a_{kj}=b_{i1}a_{1j}+b_{i2}a_{2j}+\dots+b_{in}a_{nj}$$

We see that the element of C in the ith row and jth column is equal to the sum of the products of the corresponding elements of the ith row of matrix B and the jth column of matrix A, which means it coincides with the corresponding element of the product of matrix B by matrix A. Hence, $C=BA$.

The proof is much simpler in matrix notation. Suppose

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{z}=\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

are the respective vectors. From formulas (6) and (7) we have

$$\boldsymbol{y}=A\boldsymbol{x} \quad \text{and} \quad \boldsymbol{z}=B\boldsymbol{y}$$

whence

$$\boldsymbol{z}=B(A\boldsymbol{x})=(BA)\boldsymbol{x}$$

Consequently, the matrix of the resultant transformation $C=BA$.

Example 4. Find the result of the following succession of linear transformations:

$$\begin{aligned} y_1&=5x_1-x_2+3x_3, \\ y_2&=x_1-2x_2, \\ y_3&=7x_2-x_3 \end{aligned}$$

and

$$z_1 = 2y_1 + y_3,$$
$$z_2 = y_2 - 5y_3,$$
$$z_3 = 2y_2$$

Solution. Write the transformation matrices

$$A = \begin{bmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{bmatrix}$$

Multiplying these matrices, we get the matrix of the resultant transformation

$$C = BA = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{bmatrix} = \begin{bmatrix} 10 & 5 & 5 \\ 1 & -37 & 5 \\ 2 & -4 & 0 \end{bmatrix}$$

And so the desired linear transformation has the form

$$z_1 = 10x_1 + 5x_2 - 5x_3,$$
$$z_2 = x_1 - 37x_2 + 5x_3,$$
$$z_3 = 2x_1 - 4x_2$$

10.11 INVERSE TRANSFORMATION

Suppose a given transformation

$$\left.\begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n \end{array}\right\} \qquad (1)$$

carries a set of variables $x_1, x_2, \dots, x_n$ into a set of variables $y_1, y_2, \dots, y_n$.

Definition. The transformation which carries a set of variables $y_1, y_2, \dots, y_n$ into the set of variables $x_1, x_2, \dots, x_n$ is termed an *inverse transformation* relative to the transformation (1).

We obtain the formulas of the inverse transformation by solving system (1) for the variables $x_1, x_2, \dots, x_n$.

Let A_{ij} $(i, j = 1, 2, \dots, n)$ be the cofactors of the elements a_{ij} of matrix $A = [a_{ij}]$ and

$$\det A = \det [a_{ij}] = \Delta \neq 0$$

Multiplying the equations of the system (1) by A_{11}, A_{21}, $\ldots$, A_{n1} respectively, and adding, we get, by a familiar formula,

$$A_{11}y_1 + A_{21}y_2 + \ldots + A_{n1}y_n = \Delta x_1$$

In similar fashion we derive

$$\begin{gathered} A_{12}y_1 + A_{22}y_2 + \ldots + A_{n2}y_n = \Delta x_2, \\ \cdots\cdots\cdots\cdots\cdots \\ A_{1n}y_1 + A_{2n}y_2 + \ldots + A_{nn}y_n = \Delta x_n \end{gathered}$$

whence

$$\left.\begin{aligned} x_1 &= \frac{A_{11}}{\Delta} y_1 + \frac{A_{21}}{\Delta} y_2 + \ldots + \frac{A_{n1}}{\Delta} y_n, \\ x_2 &= \frac{A_{12}}{\Delta} y_1 + \frac{A_{22}}{\Delta} y_2 + \ldots + \frac{A_{n2}}{\Delta} y_n, \\ &\cdots\cdots\cdots\cdots\cdots \\ x_n &= \frac{A_{1n}}{\Delta} y_1 + \frac{A_{2n}}{\Delta} y_2 + \ldots + \frac{A_{nn}}{\Delta} y_n \end{aligned}\right\} \tag{2}$$

Thus, the inverse transformation of a linear transformation (if it exists) is also linear.

Theorem. *A linear transformation has a unique inverse if and only if the matrix of the given transformation is nonsingular. The inverse of a linear transformation is linear and its matrix is the inverse of the matrix of the initial transformation.*

Proof. If $A = [a_{ij}]$ is the matrix of transformation (1) and $\Delta = \det A \neq 0$, then the inverse exists and is defined by formulas (2). The matrix of the inverse transformation is clearly equal to

$$\left(\frac{A_{ji}}{\Delta}\right) = A^{-1}$$

If $\Delta = 0$, then, by routine algebra, the equations (1) cannot be uniquely solved for the variables x_1, x_2, $\ldots$, x_n. Thus, a unique inverse transformation does not exist, and there will definitely be values of the variables y_1, y_2, $\ldots$, y_n for which there are no corresponding values of the variables x_1, x_2, $\ldots$, x_n. In this case, the linear transformation is called *singular (degenerate)*.

Note 1. Write the transformation (1) in matrix form

$$\boldsymbol{y} = A\boldsymbol{x} \tag{3}$$

where $A = [a_{ij}]$ is the transformation matrix and $\boldsymbol{x}$ and $\boldsymbol{y}$ are column vectors.

If the transformation A is *nonsingular* ($\det A \neq 0$), then the inverse transformation

$$\boldsymbol{x} = A^{-1}\boldsymbol{y} \tag{4}$$

exists and each vector $\boldsymbol{x}$ of the n-dimensional space $Ox_1x_2 \ldots x_n$ is associated, by virtue of formula (3), with one and only one vector $\boldsymbol{y}$ of that space; thus, formula (3) transforms the space $Ox_1x_2 \ldots \ldots x_n$ into itself.

If the transformation A is *singular* ($\det A = 0$), then (3) transforms the space $Ox_1x_2 \ldots x_n$ into a subspace of a smaller number of dimensions.

Example. Consider the projection transformation (see Sec. 10.10, Example 2) defined by the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Here A is singular and the transformation $\boldsymbol{y} = A\boldsymbol{x}$ carries the space Ox_1x_2 into the coordinate axis Ox_1.

Note 2. We take $E\boldsymbol{x}$ to mean the identical transformation that leaves the vector $\boldsymbol{x}$ unchanged.

Since the relations

$$\boldsymbol{y} = A\boldsymbol{x} \quad \text{and} \quad \boldsymbol{x} = A^{-1}\boldsymbol{y}$$

imply

$$\boldsymbol{y} = AA^{-1}\boldsymbol{y} \quad \text{and} \quad \boldsymbol{x} = A^{-1}A\boldsymbol{x}$$

it follows that

$$AA^{-1} = A^{-1}A = E$$

10.12 EIGENVECTORS AND EIGENVALUES OF A MATRIX

Given a square matrix $A = [a_{ij}]$. Consider the linear transformation

$$\boldsymbol{y} = A\boldsymbol{x} \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are n-dimensional vectors (column matrices) of, generally speaking, a complex n-dimensional space.

Definition 1. A nonzero vector is called an *eigenvector* of a given matrix (or of the linear transformation defined by it) if as a result of an appropriate linear transformation the vector is carried into a collinear vector; that is, if the transformed vector differs from the original one only by a scalar factor.

In other words, the vector $\boldsymbol{x} \neq \boldsymbol{0}$ is an *eigenvector* of matrix A if the matrix carries $\boldsymbol{x}$ into the vector

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{2}$$

The scalar λ in (2) is called an *eigenvalue* (or *characteristic root*, *characteristic number*, *latent root*) of the matrix A which eigenvalue corresponds to the given eigenvector $\boldsymbol{x}$.

Example 1. Let us consider the projection transformation in two-dimensional space Ox_1x_2 defined by the matrix

$$A=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$$

Here the eigenvectors are (1) the nonzero vectors $\boldsymbol{x}$ directed along the x_1-axis with eigenvalue $\lambda_1=1$ and (2) the nonzero vectors $\boldsymbol{y}$ directed along the $\boldsymbol{x}_2$-axis with eigenvalue $\lambda_2=0$ (Fig. 54).

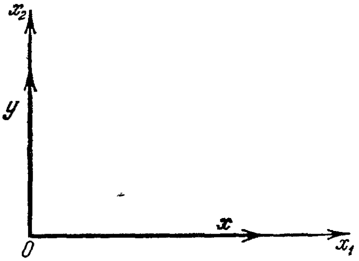


Fig. 54

Theorem 1. *Every linear transformation (matrix) in a complex vector space has at least one real or complex eigenvector.*

Proof. Let A be the matrix of a linear transformation. The eigenvectors of A are nonzero solutions of the matrix equation

$$A\boldsymbol{x}=\lambda\boldsymbol{x}$$

or

$$(A-\lambda E)\,\boldsymbol{x}=0 \qquad (3)$$

where the matrix $A-\lambda E$ is called the *characteristic matrix*. Equation (3) is a homogeneous linear system which has nontrivial solutions if and only if the determinant of the system is zero; thus, the following condition must be met:

$$\det\,(A-\lambda E)=0 \qquad (4)$$

The determinant (4) is called the *characteristic* (*secular*) *determinant* of matrix A, and equation (4) is called the *characteristic* (*secular*) *equation* of matrix A. Expanded, the characteristic equation (4) looks like this

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22}-\lambda & \dots & a_{2n}\\ \dots & \dots & \dots & \dots\\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix}=0 \qquad (4')$$

or

$$\lambda^n-\sigma_1\lambda^{n-1}+\sigma_2\lambda^{n-2}-\dots+(-1)^{n-1}\sigma_{n-1}\lambda+(-1)^n\sigma_n=0 \qquad (5)$$

The polynomial in the left member of (5) is called the *characteristic polynomial* of matrix A. Its coefficients $\sigma_i\,(i=1, 2, \dots, n)$ are defined by the following rules. The coefficient σ_1 is equal to the sum of the diagonal elements of matrix A, or $\sigma_1=\sum\limits_{i=1}^{n}a_{ii}$. This number is called the *trace of matrix* A and is denoted by $\sigma_1 = \operatorname{tr} A$. The coefficient σ_2 is the sum of all diagonal second-order minors of A. Generally, the coefficient σ_k is the sum of all the diagonal kth-order minors of A. Finally, the constant term σ_n is equal to the determinant of the matrix A:

$$\sigma_n = \det A$$

The characteristic equation (5) is an algebraic equation of degree n in λ, and, hence, as is proved in algebra, has at least one real or complex root. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ $(m \leqslant n)$ be distinct roots of equation (5). These roots are called *eigenvalues* (or *characteristic numbers*) of the matrix A, and the set of *all* eigenvalues is termed the *spectrum of the matrix* A. Take some root $\lambda = \lambda_j$ and substitute it into (4) to get $(A - \lambda_j E)\, \boldsymbol{x} = \mathbf{0}$ or, in expanded form,

$$\left.\begin{array}{l} (a_{11} - \lambda_j)\, x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0, \\ a_{21}x_1 + (a_{22} - \lambda_j)\, x_2 + \ldots + a_{2n}x_n = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn} - \lambda_j)\, x_n = 0 \end{array}\right\} \qquad (6)$$

Since the determinant of system (6) $\det (A - \lambda_j E) = 0$, this system definitely has nontrivial solutions, which are the eigenvectors of A that correspond to the eigenvalue λ_j. If the rank of the matrix $A - \lambda_j E$ is equal to r $(r < n)$, then there are $k = n - r$ linearly independent eigenvectors

$$\boldsymbol{x}^{(1j)}, \ \boldsymbol{x}^{(2j)}, \ \ldots, \ \boldsymbol{x}^{(kj)}$$

associated with the root λ_j. The theorem is proved.

Note. It may be proved that the number of linearly independent eigenvectors corresponding to one and the same root of the characteristic equation does not exceed the multiplicity of that root. One thing that this implies is that if the roots of the characteristic equation (5) are distinct, then each eigenvalue is associated with one and only one eigenvector to within the factor of proportionality.

Example 2. Find the eigenvalues and the eigenvectors of the matrix

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Solution. Set up the characteristic equation of the matrix A:

$$\begin{bmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{bmatrix} = 0$$

whence $(\lambda-1)^2(4-\lambda)=0$ and $\lambda_1=\lambda_2=1$, $\lambda_3=4$. Take $\lambda_1=1$ and put it into equation

$$(A-\lambda_j E)\,\boldsymbol{x}=\mathbf{0} \tag{7}$$

to get

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\mathbf{0}$$

or

$$\left.\begin{aligned} x_1+x_2+x_3&=0, \\ x_1+x_2+x_3&=0, \\ x_1+x_2+x_3&=0 \end{aligned}\right\} \tag{8}$$

Since the rank of the matrix of system (8) $r=1$, two of the equations are consequences of the third (which, incidentally, is obvious). It therefore suffices to solve the equation

$$x_1+x_2+x_3=0$$

Putting $x_1=c_1$, $x_2=c_2$, we get

$$x_3=-(c_1+c_2)$$

where c_1 and c_2 are arbitrary scalars not simultaneously zero.

In particular, first choosing $c_1=1$, $c_2=0$ and then $c_1=0$, $c_2=1$, we will have the simplest fundamental set of solutions consisting of two linearly independent eigenvectors of matrix A:

$$\boldsymbol{x}^{(1)}=\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}^{(2)}=\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

All the other eigenvectors of A that correspond to the eigenvalue $\lambda_1=1$ are linear combinations of these basis vectors and fill the plane spanned by the vectors $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ (with the exception of the origin).

Now take $\lambda_3=4$. Substituting this value into (7), we get

$$\begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\mathbf{0}$$

or

$$\left.\begin{aligned} -2x_1+x_2+x_3&=0, \\ x_1-2x_2+x_3&=0, \\ x_1+x_2-2x_3&=0 \end{aligned}\right\} \tag{9}$$

The rank of the system (9) is $r=2$, and the upper left minor

$$\delta=\begin{vmatrix} -2 & 1 \\ 1 & -2 \end{vmatrix} \neq 0$$

Thus, the third equation of the system is a consequence of the first two equations, and so we can confine ourselves to a system of the first two equations:

$$\left.\begin{aligned} -2x_1+x_2+x_3=0, \\ x_1-2x_2+x_3=0 \end{aligned}\right\}$$

whence

$$\frac{x_1}{\begin{vmatrix} 1 & 1 \\ -2 & 1 \end{vmatrix}}=\frac{x_2}{-\begin{vmatrix} -2 & 1 \\ 1 & 1 \end{vmatrix}}=\frac{x_3}{\begin{vmatrix} -2 & 1 \\ 1 & -2 \end{vmatrix}}$$

or

$$\frac{x_1}{3}=\frac{x_2}{3}=\frac{x_3}{3}; \quad \text{that is,} \quad x_1=x_2=x_3=c$$

where c is a constant different from zero.

Putting $c=1$, we get the simplest solution that effects the eigenvector of A

$$\boldsymbol{x}^{(3)}=\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Definition 2. A linear subspace E_k $(k \leqslant n)$ is said to be *invariant* with respect to a given linear transformation

$$\boldsymbol{y}=A\boldsymbol{x}$$

if each transformed vector of this subspace also belongs to it; thus, $\boldsymbol{x} \in E_k$ implies $A\boldsymbol{x} \in E_k$.

The proof of Theorem 1 clearly holds if we consider, in some invariant space, a linear transformation defined by matrix A.

Theorem 1′. *Every linear transformation defined on an invariant subspace of a complex vector space has at least one eigenvector.*

We note one more important property of eigenvectors.

Theorem 2. *The eigenvectors of a matrix which correspond to pairwise distinct eigenvalues are linearly independent.*

Proof. Let A be a given matrix and

$$A\boldsymbol{x}^{(j)}=\lambda_j\boldsymbol{x}^{(j)} \qquad (j=1, 2, \ldots, m) \tag{10}$$

where

$$\boldsymbol{x}^{(j)} \neq \boldsymbol{0} \quad \text{and} \quad \lambda_j \neq \lambda_k \quad \text{for} \quad j \neq k$$

Suppose that

$$c_1\boldsymbol{x}^{(1)}+c_2\boldsymbol{x}^{(2)}+\ldots+c_m\boldsymbol{x}^{(m)}=\boldsymbol{0} \tag{11}$$

where $|c_1|+|c_2|+\ldots+|c_m|\neq 0$.

For the sake of definiteness, let $c_1\neq 0$. Applying the transformation A to (11), we get, by virtue of formulas (10),

$$\lambda_1 c_1\boldsymbol{x}^{(1)}+\lambda_2 c_2\boldsymbol{x}^{(2)}+\ldots+\lambda_m c_m\boldsymbol{x}^{(m)}=\boldsymbol{0} \tag{12}$$

Now, multiplying (11) by λ_m and subtracting (12) from the resulting equation, we find

$$(\lambda_m-\lambda_1)\,c_1\boldsymbol{x}^{(1)}+(\lambda_m-\lambda_2)\,c_2\boldsymbol{x}^{(2)}+\ldots+(\lambda_m-\lambda_{m-1})\,c_{m-1}\boldsymbol{x}^{(m-1)}=\boldsymbol{0} \tag{13}$$

Then, from (13) we can eliminate the vector $\boldsymbol{x}^{(m-1)}$ in similar fashion, and so on. Thus, eliminating the vectors

$$\boldsymbol{x}^{(m)},\ \boldsymbol{x}^{(m-1)},\ \ldots,\ \boldsymbol{x}^{(2)}$$

we get

$$(\lambda_m-\lambda_1)(\lambda_{m-1}-\lambda_1)\ \ldots\ (\lambda_2-\lambda_1)\,c_1\boldsymbol{x}^{(1)}=\boldsymbol{0} \tag{14}$$

But this equation is impossible because not one of the factors of the left-hand member is equal to zero. Thus, our assumption is false and the eigenvectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ are linearly independent.

Corollary. If all the eigenvalues of a matrix A of order n are pairwise distinct, then the corresponding eigenvectors of this matrix (n altogether[1]) form a basis of an appropriate n-dimensional space.

10.13 SIMILAR MATRICES

Definition. Two matrices associated with one and the same linear transformation in different bases are called *similar*.

If a matrix A is similar to a matrix B, we write $A \backsim B$. Let us derive a condition for the similarity of two matrices. Suppose, in a certain basis, matrix A represents the linear transformation

$$\boldsymbol{y}=A\boldsymbol{x} \tag{1}$$

In a new basis (in new coordinates), this same linear transformation will be described by another matrix, B:

$$\eta=B\xi \tag{2}$$

where

$$B \backsim A$$

---

[1] It is assumed that one eigenvector is taken for each eigenvalue.

Denote by S the change-of-basis matrix from the new system to the old system; thus, let

$$x = S\xi, \qquad y = S\eta \tag{3}$$

where

$$\det S \neq 0$$

Putting formulas (3) into (1), we get

$$S\eta = AS\xi$$

Then, premultiplying the last equation by the inverse matrix S^{-1}, we get

$$\eta = S^{-1}AS\xi \tag{4}$$

Comparing formulas (4) and (2), we obtain

$$B = S^{-1}AS \tag{5}$$

With regard to the matrices A and B connected by the relation (5), we say that the matrix B is obtained from A *via a transformation by means of matrix* S. We thus conclude that *two matrices are similar if and only if one of them is obtained from the other by a transformation effected by some nonsingular matrix.*

From (5) we derive $A = SBS^{-1}$, which is to say, *if matrix* B *is similar to matrix* A, *then also, conversely, matrix* A *is similar to matrix* B. The following are some properties of a transformation by means of a matrix S.

1. *The transformation of a sum is equal to the sum of the transformations:*

$$S^{-1}(A+B)S = S^{-1}AS + S^{-1}BS$$

2. *The transformation of a product is equal to the product of the transformations of the factors:*

$$S^{-1}(AB)S = S^{-1}AS \cdot S^{-1}BS$$

3. *The transformation of an inverse matrix is equal to the inverse of the transformed matrix:*

$$S^{-1}A^{-1}S = (S^{-1}AS)^{-1}$$

4. *The transformation of an integral power (positive or negative) is equal to the same power of the transformation:*

$$S^{-1}A^nS = (S^{-1}AS)^n$$

Theorem 1. *Similar matrices have the same characteristic polynomials.*

Proof. Let $B \backsim A$. It is required to prove that

$$\det(A - \lambda E) = \det(B - \lambda E)$$

Since

$$B = S^{-1}AS \quad (\det S \neq 0)$$

then

$$\det(B - \lambda E) = \det[S^{-1}(A - \lambda E)S] = \\ = \det S^{-1} \det(A - \lambda E) \det S = \det(A - \lambda E)\,^{1)}$$

Thus

$$\det(B - \lambda E) = \det(A - \lambda E)$$

Corollary 1. Similar matrices have identical traces and the same eigenvalues (including their multiplicities).

Corollary 2. The property of a vector to be an eigenvector of a given linear transformation is independent of the choice of basis. Indeed, suppose

$$Ax = \lambda x \qquad (x \neq 0)$$

If in the new basis the vector x is equivalent to the vector ξ, then we have

$$x = S\xi$$

where S is the change-of-basis matrix.

From this we have $AS\xi = \lambda S\xi$ and, hence, $S^{-1}AS\xi = \lambda\xi$, which means that ξ is an eigenvector of the matrix $B = S^{-1}AS \backsim A$ that describes our linear transformation in the new basis.

Note. Since the characteristic polynomial, the eigenvalues and the eigenvectors are the same for all matrices representing a given linear transformation, they are called, respectively, the *characteristic polynomial, eigenvalues and eigenvectors of the linear transformation itself.*

Theorem 2. *If a given square matrix of order* n *has* n *linearly independent eigenvectors, then, taking them for basis vectors, we obtain a diagonal matrix similar to the given one.*

Proof. Suppose we have a square matrix A. Form a basis from the eigenvectors $e_1, e_2, \ldots, e_n$. Since e_j are eigenvectors, it follows that

$$Ae_j = \lambda_j e_j \qquad (j = 1, 2, \ldots, n)$$

---

[1] Here we make use of familiar theorems (see Secs. 7.2 and 7.4): (1) the determinant of the product of two square matrices of the same order is equal to the product of the determinants of the matrices; (2) the determinant of an inverse matrix is equal to the reciprocal of the determinant of the original matrix.

Consider any vector $\boldsymbol{x}$ of our space. Decomposing it into its basis vectors $\boldsymbol{e}_j$ $(j=1,\ 2,\ \ldots,\ n)$, we have

$$\boldsymbol{x}=\sum_{j=1}^{\infty} x_j\boldsymbol{e}_j$$

where x_j are the coordinates of the vector $\boldsymbol{x}$ in the given basis.

Applying the transformation A to the vector $\boldsymbol{x}$, we get a new vector

$$\boldsymbol{y}=A\boldsymbol{x}=A\sum_{j=1}^{n} x_j\boldsymbol{e}_j$$

or, since the transformation A is linear,

$$\boldsymbol{y}=\sum_{j=1}^{n} x_jA\boldsymbol{e}_j=\sum_{j=1}^{n} x_j\lambda_j\boldsymbol{e}_j$$

whence we see that the coordinates of the vector $\boldsymbol{y}$ in the given basis are

$$y_j=\lambda_jx_j \qquad (j=1,\ 2,\ \ldots,\ n)$$

or

$$y_j=\sum_{k=1}^{n} \delta_{jk}\lambda_jx_k$$

where δ_{jk} is the Kronecker delta.

Thus, in the new basis the transformation matrix is the diagonal matrix

$$\Lambda=(\delta_{jk}\lambda_j)$$

or, expanded,

$$\Lambda=\begin{bmatrix}\lambda_1 & 0 & 0 \ldots 0\\ 0 & \lambda_2 & 0 \ldots 0\\ . & . & . \ . \ . \ . \ .\\ 0 & 0 & 0 \ldots \lambda_n\end{bmatrix}$$

Corollary. Any square matrix whose eigenvalues are pairwise distinct can be reduced to diagonal form by means of a similarity transformation.

This result follows directly from Theorem 2 of the preceding section.

10.14 BILINEAR FORM OF A MATRIX

Let $A=[a_{jk}]$ be a real square matrix and let $\boldsymbol{x}$, $\boldsymbol{y}$ be vectors in an n-dimensional complex space. Form the scalar product

$$(A\boldsymbol{x}, \boldsymbol{y})=\sum_{j=1}^{n}(A\boldsymbol{x})_j y_j^* = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_k y_j^* \tag{1}$$

Expression (1) is called a *bilinear form of matrix A*.

Let us derive an important property of a bilinear form. The sum (1) will plainly have the same value if we change the order of summation and at the same time interchange the summation indices. We therefore get

$$(A\boldsymbol{x}, \boldsymbol{y})=\sum_{j=1}^{n}\sum_{k=1}^{n} a_{kj}x_j y_k^*$$

Let us write this sum as a scalar product:

$$(A\boldsymbol{x}, \boldsymbol{y})=\sum_{j=1}^{n}\sum_{k=1}^{n} a_{kj}x_j y_k^* = \left(\sum_{j=1}^{n}\sum_{k=1}^{n} a_{kj}y_k x_j^*\right)^* = (A'\boldsymbol{y}, \boldsymbol{x})^* = (\boldsymbol{x}, A'\boldsymbol{y})$$

Thus,

$$(A\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{x}, A'\boldsymbol{y}) \tag{2}$$

This means that *in the scalar product* (1) *the real matrix* A *may be moved from first to second position if we replace it by its transform.*

Corollary. If matrix A is real and symmetric ($A'=A$), then

$$(Ax, y)=(x, Ay) \tag{3}$$

In a scalar product, a real symmetric matrix may be moved from first position to second.

10.15 PROPERTIES OF SYMMETRIC MATRICES

Theorem 1. *All the eigenvalues of a symmetric matrix with real elements are real.*

Proof. Let λ be an eigenvalue of matrix A and $\boldsymbol{x}$ the corresponding eigenvector; thus,

$$A\boldsymbol{x}=\lambda\boldsymbol{x} \qquad (\boldsymbol{x}\neq 0) \tag{1}$$

Since $A'=A$, then

$$(A\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{x}, A\boldsymbol{x})$$

or, by (1),

$$(\lambda\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{x}, \lambda\boldsymbol{x})$$

whence

$$\lambda(\boldsymbol{x},\ \boldsymbol{x})=\lambda^*(\boldsymbol{x},\ \boldsymbol{x})$$

The eigenvector is nonzero by definition, and so

$$(\boldsymbol{x},\ \boldsymbol{x})\neq 0$$

and, consequently, $\lambda=\lambda^*$, or λ is a real number.

Corollary. The characteristic equation of a symmetric real matrix has only real roots.

Theorem 2. *The eigenvectors of a symmetric real matrix that correspond to distinct eigenvalues are orthogonal among themselves.*

Proof. Let A be a symmetric real matrix. Consider two eigenvectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ associated with the eigenvalues λ_i and λ_j $(\lambda_i\neq\lambda_j)$. We have

$$A\boldsymbol{x}^{(i)}=\lambda_i\boldsymbol{x}^{(i)} \tag{2}$$

and

$$A\boldsymbol{x}^{(j)}=\lambda_j\boldsymbol{x}^{(j)} \tag{3}$$

Form the scalar product

$$(A\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)})=(\boldsymbol{x}^{(i)},\ A\boldsymbol{x}^{(j)})$$

From this, by (2) and (3), we get

$$(\lambda_i\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)})=(\boldsymbol{x}^{(i)},\ \lambda_j\boldsymbol{x}^{(j)})$$

and

$$\lambda_i(\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)})=\lambda_j^*(\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)}) \tag{4}$$

Since on the basis of Theorem 1 the eigenvalue λ_j is real, it follows that $\lambda_j^*=\lambda_j$.

Hence, from (4), we have

$$(\lambda_i-\lambda_j)(\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)})=0$$

But

$$\lambda_i-\lambda_j\neq 0$$

and so

$$(\boldsymbol{x}^{(i)},\ \boldsymbol{x}^{(j)})=0$$

or the eigenvectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are orthogonal among themselves.

Note. The eigenvectors of a symmetric matrix with real elements may be assumed to be real.

Theorem 3. *Any real symmetric matrix can be reduced to diagonal form by means of a similarity transformation.*

Proof. For the purpose of pictorialness, let us confine our proof to three-dimensional space, E_3.

Suppose we have a symmetric matrix A of order three. As we know, every matrix has at least one eigenvector (see Sec. 10.12, Theorem 1). Denote by e_1 an eigenvector of matrix A. Since A is symmetric, the eigenvector e_1 can be chosen real.

Consider all eigenvectors x that are orthogonal to e_1, that is such that

$$(x, e_1) = 0 \tag{5}$$

We will show that they form an invariant subspace E_2 with respect to the transformation A (Fig. 55).

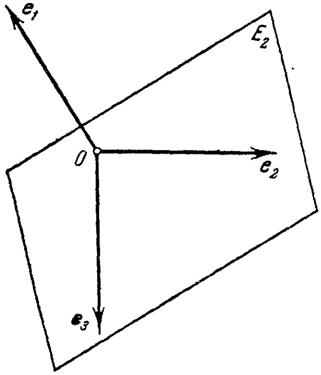


Fig. 55

First of all, if $x \in E_2$ and $y \in E_2$, that is,

$$(x, e_1) = (y, e_1) = 0$$

then for arbitrary scalars α and β we have

$$(\alpha x + \beta y, e_1) = \alpha (x, e_1) + \beta (y, e_1) = 0$$

and, consequently,

$$\alpha x + \beta y \in E_2$$

Thus, the set of eigenvectors satisfying Condition (5) form a linear space, and it is easy to see that this space is two-dimensional.

Now let $x \in E_2$. Consider the scalar product

$$(Ax, e_1) = (x, Ae_1) = (x, \lambda_1 e_1) = \lambda_1 (x, e_1) = 0$$

Thus,

$$Ax \in E_2$$

By Theorem 1′ (see Sec. 10.12), there also exists in the two-dimensional space E_2 an eigenvector e_2 of matrix A. Now consider the eigenvectors x that are orthogonal both to e_1 and to e_2, that is, such that

$$(x, e_1) = (x, e_2) = 0$$

Similarly it is proved that these eigenvectors form a one-dimensional space E_1 that is invariant under the transformation A. Again, space E_1 has an eigenvector e_3 of matrix A. The eigenvectors e_1, e_2, e_3 are pairwise orthogonal and so are linearly independent. We thus construct an orthogonal basis for the space E_3 consisting of the eigenvectors of the matrix A.

Denote by λ_j the eigenvalues associated with the eigenvectors e_j. By Theorem 2 of Sec. 10.13, the matrix Λ of the given linear transformation will be diagonal in the proper basis e_1, e_2, e_3, and in our case,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

The proof of the theorem in the general case is analogous.

Corollary 1. For every linear transformation with a real symmetric matrix there is an orthogonal basis (consisting of real eigenvectors of the given matrix) in which the transformation matrix is diagonal.

Corollary 2. If the matrix is symmetric, then every eigenvalue is associated with as many linearly independent eigenvectors as is the multiplicity of the eigenvalue.

Theorem 4 (extremal property of eigenvalues). *Let* A *be a real symmetric matrix and*

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_n),$$
$$\Lambda = \max(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ *are all the eigenvalues of* A.

Then the following inequality is valid for any vector **x**:

$$\lambda(\boldsymbol{x}, \boldsymbol{x}) \leqslant (A\boldsymbol{x}, \boldsymbol{x}) \leqslant \Lambda(\boldsymbol{x}, \boldsymbol{x}) \tag{6}$$

Proof. By virtue of Corollary 1 of Theorem 3, there is a set of eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of the matrix A:

$$A\boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j \qquad (j = 1, 2, \ldots, n)$$

which form an orthonormal basis of the space E_n. Then any vector $\boldsymbol{x}$ can be represented in the form

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \ldots + x_n\boldsymbol{e}_n$$

where $x_1, x_2, \ldots, x_n$ are the coordinates of vector $\boldsymbol{x}$ in the given basis, whence

$$A\boldsymbol{x} = x_1A\boldsymbol{e}_1 + x_2A\boldsymbol{e}_2 + \ldots + x_nA\boldsymbol{e}_n = \lambda_1x_1\boldsymbol{e}_1 + \lambda_2x_2\boldsymbol{e}_2 + \ldots + \lambda_nx_n\boldsymbol{e}_n$$

Taking into account the orthogonality of the vectors of the basis, we have

$$(A\boldsymbol{x}, \boldsymbol{x}) = \left(\sum_{j=1}^{n} \lambda_j x_j \boldsymbol{e}_j, \sum_{k=1}^{n} x_k \boldsymbol{e}_k\right) =$$
$$= \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j x_j x_k^* (\boldsymbol{e}_j, \boldsymbol{e}_k) = \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j x_j x_k^* \delta_{jk} = \sum_{j=1}^{n} \lambda_j |x_j|^2$$

or

$$(A\boldsymbol{x}, \boldsymbol{x}) = \sum_{j=1}^{n} \lambda_j |x_j|^2 \tag{7}$$

Replacing λ_j by the least value of λ in (7), we obtain

$$(A\boldsymbol{x}, \boldsymbol{x}) \geqslant \lambda \sum_{j=1}^{n} |x_j|^2 = \lambda(\boldsymbol{x}, \boldsymbol{x})$$

Similarly, putting in (7) the greatest value of Λ in place of λ_j, we find

$$(A\boldsymbol{x}, \boldsymbol{x}) \leqslant \Lambda \sum_{j=1}^{n} |x_j|^2 = \Lambda(\boldsymbol{x}, \boldsymbol{x})$$

And inequality (6) is proved.

Corollary. The minimal eigenvalue λ and the maximal eigenvalue Λ of the symmetric real matrix A are, respectively, the smallest and largest values of the quadratic form

$$u = (A\boldsymbol{x}, \boldsymbol{x})$$

on a unit sphere $(\boldsymbol{x}, \boldsymbol{x}) = 1$.

Indeed, putting $(\boldsymbol{x}, \boldsymbol{x}) = 1$ in inequality (6), we have

$$\lambda \leqslant (A\boldsymbol{x}, \boldsymbol{x}) \leqslant \Lambda$$

Moreover, if $A\boldsymbol{x} = \lambda\boldsymbol{x}$, then

$$(A\boldsymbol{x}, \boldsymbol{x}) = (\lambda\boldsymbol{x}, \boldsymbol{x}) = \lambda$$

Similarly, if $A\boldsymbol{x} = \Lambda\boldsymbol{x}$, then

$$(A\boldsymbol{x}, \boldsymbol{x}) = (\Lambda\boldsymbol{x}, \boldsymbol{x}) = \Lambda$$

Thus,

$$\lambda = \min (A\boldsymbol{x}, \boldsymbol{x}) \quad \text{for} \quad (\boldsymbol{x}, \boldsymbol{x}) = 1$$

and

$$\Lambda = \max (A\boldsymbol{x}, \boldsymbol{x}) \quad \text{for} \quad (\boldsymbol{x}, \boldsymbol{x}) = 1$$

A symmetric real matrix $A = [a_{ij}]$ will be called *positive definite* if the corresponding quadratic form

$$u = (A\boldsymbol{x}, \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j^*$$

is positive definite (see Sec. 8.13), that is, for any vector $\boldsymbol{x} \neq \boldsymbol{0}$ we have

$$(A\boldsymbol{x}, \boldsymbol{x}) > 0$$

Theorem 5. *A symmetric real matrix is positive definite if and only if all its eigenvalues are positive.*

Proof. If A is a symmetric real matrix and its eigenvalues λ_j are such that $\lambda_j > 0$ $(j = 1, 2, \ldots, n)$, then on the basis of formula (7) we have, from the proof of the preceding theorem,

$$(A\boldsymbol{x}, \boldsymbol{x}) = \sum_{j=1}^{n} \lambda_j |x_j|^2$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, whence for $\boldsymbol{x} \neq \mathbf{0}$ we have

$$(A\boldsymbol{x}, \boldsymbol{x}) > 0$$

Thus, the matrix A is positive definite.

Conversely, let A be a positive definite symmetric real matrix. By Theorem 1, all its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real, and

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

is the smallest value of the quadratic form $u = (A\boldsymbol{x}, \boldsymbol{x})$ on the sphere $(\boldsymbol{x}, \boldsymbol{x}) = 1$. Since the sphere $(\boldsymbol{x}, \boldsymbol{x}) = 1$ is a compact bounded set and the quadratic form u is continuous and positive on this sphere, then by the Weierstrass theorem, there is a smallest value of u and it is positive; that is,

$$\lambda > 0$$

whence, all the more so,

$$\lambda_j > 0 \quad \text{for} \quad j = 1, 2, \ldots, n$$

We give without proof the conditions of a positive definite real matrix [2].

Theorem 6. *For a real matrix* $A = [a_{ij}]$, *where* $a_{ij} = a_{ji}$, *to be positive definite, it is necessary and sufficient that the following Sylvester conditions be fulfilled:*

$$\Delta_1 = a_{11} > 0, \qquad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \ \ldots,$$

$$\Delta_n = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} > 0$$

Thus, a symmetric real matrix A is positive definite if and only if all the principal diagonal minors of its determinant det A are strictly positive.

*10.16 PROPERTIES OF MATRICES WITH REAL ELEMENTS

In the sequel we will, as a rule, consider matrices $A = [a_{ij}]$ whose elements a_{ij} are real. These are called *real matrices*.

Let $A = [a_{ij}]$ be a real square matrix of order n. Since its characteristic equation

$$\det(A-\lambda E)=0$$

is a polynomial with real coefficients, the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the characteristic equation, which are the eigenvalues of the matrix A, are conjugate in pairs if they are complex (see Sec. 5.1); that is, if λ_s is an eigenvalue of A, then the conjugate λ_s^* is also an eigenvalue of the matrix A and is of the same multiplicity.

A real matrix may not have real eigenvalues. However, in one important case when the elements of the matrix are positive, the existence of at least one real eigenvalue is guaranteed [6].

Perron's theorem. *If all the elements of a square matrix are positive, then the numerically largest eigenvalue is also positive and is a simple root of the characteristic equation of the matrix; it is associated with an eigenvector with positive coordinates.*

Eigenvectors of a real matrix A with distinct eigenvalues are complex in the general case and do not possess the property of orthogonality. However, by invoking the eigenvectors of the transpose A' we can obtain the so-called *biorthogonality relations*, which for the case of a symmetric matrix are equivalent to the ordinary relations of orthogonality.

Theorem 1. *If a matrix* A *is real and its eigenvalues are pairwise distinct, then there exist two bases,* $\{x_j\}$ *and* $\{x'_j\}$, *of the space* E_n *consisting respectively of the eigenvectors of matrix* A *and the eigenvectors of the transpose* A′ *which satisfy the following conditions of biorthonormalization:*

$$(x_j, x'_k)=\begin{cases}0 & \text{for } j\neq k,\\ 1 & \text{for } j=k\end{cases}$$

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the matrix A. Since A is real, we know that its eigenvalues are conjugate in pairs, which means that besides the eigenvalue λ_j the conjugate λ_j^* is also an eigenvalue of A. Denote by x_j $(j=1, 2, \ldots, n)$ the corresponding eigenvectors of matrix A, that is,

$$Ax_j=\lambda_j x_j \qquad (j=1, 2, \ldots, n) \tag{1}$$

The eigenvectors $\{x_j\}$ form a basis for the space E_n (Sec. 10.12, Theorem 2, Corollary).

Since a determinant remains unaltered under an interchange of rows and columns,

$$\det(A'-\lambda E)\equiv\det(A-\lambda E)$$

and, consequently, the transpose A' has the same eigenvalues λ_j as the matrix A. Let x'_j $(j=1, 2, \ldots, n)$ be the eigenvectors of the matrix A' which correspond to the conjugate eigenvalues λ_j^*, that is,

$$A'\boldsymbol{x}_j' = \lambda_j^* \boldsymbol{x}_j' \qquad (j = 1, 2, \ldots, n) \tag{2}$$

The eigenvectors $\{\boldsymbol{x}_j'\}$ also form a basis for the space E_n.

The bases $\{\boldsymbol{x}_j\}$ and $\{\boldsymbol{x}_j'\}$ are *biorthogonal*, namely:

$$(\boldsymbol{x}_j, \boldsymbol{x}_k') = 0 \quad \text{for} \quad j \neq k \tag{3}$$

Indeed, on the one hand, we have

$$(A\boldsymbol{x}_j, \boldsymbol{x}_k') = (\lambda_j \boldsymbol{x}_j, \boldsymbol{x}_k') = \lambda_j (\boldsymbol{x}_j, \boldsymbol{x}_k') \tag{4}$$

On the other, taking into consideration that A is real, we get

$$(A\boldsymbol{x}_j, \boldsymbol{x}_k') = (\boldsymbol{x}_j, A'\boldsymbol{x}_k') = (\boldsymbol{x}_j, \lambda_k^* \boldsymbol{x}_k') = \lambda_k (\boldsymbol{x}_j, \boldsymbol{x}_k') \tag{5}$$

From (4) and (5) we derive

$$\lambda_j (\boldsymbol{x}_j, \boldsymbol{x}_k') = \lambda_k (\boldsymbol{x}_j, \boldsymbol{x}_k') \tag{6}$$

Since $\lambda_j \neq \lambda_k$ when $j \neq k$, from (6) follows (3). We will show that the eigenvectors $\{\boldsymbol{x}_j\}$ and $\{\boldsymbol{x}_j'\}$ may be normalized so that

$$(\boldsymbol{x}_j, \boldsymbol{x}_j') = 1 \qquad (j = 1, 2, \ldots, n) \tag{7}$$

Indeed, decomposing $\boldsymbol{x}_j$ in terms of the eigenvectors of the basis $\{\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n'\}$ we have

$$\boldsymbol{x}_j = c_1 \boldsymbol{x}_1' + \ldots + c_j \boldsymbol{x}_j' + \ldots + c_n \boldsymbol{x}_n'$$

whence, taking into account the biorthogonality condition (3), we get

$$(\boldsymbol{x}_j, \boldsymbol{x}_j') = c_1^* (\boldsymbol{x}_j, \boldsymbol{x}_1') + \ldots + c_j^* (\boldsymbol{x}_j, \boldsymbol{x}_j') + \ldots$$
$$\ldots + c_n^* (\boldsymbol{x}_j, \boldsymbol{x}_n') = c_j^* (\boldsymbol{x}_j, \boldsymbol{x}_j') > 0$$

and so

$$(\boldsymbol{x}_j, \boldsymbol{x}_j') = \alpha_j \neq 0$$

Taking the eigenvectors $\frac{1}{\alpha_1^*}\boldsymbol{x}_1', \ldots, \frac{1}{\alpha_n^*}\boldsymbol{x}_n'$ in place of $\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'$, we get the required normalization (7) since

$$\left(\boldsymbol{x}_j, \frac{1}{\alpha_j^*}\boldsymbol{x}_j'\right) = \frac{1}{\alpha_j}(\boldsymbol{x}_j, \boldsymbol{x}_j') = \frac{1}{\alpha_j} \cdot \alpha_j = 1 \qquad (j = 1, 2, \ldots, n)$$

Thus, if the eigenvalues of the real matrix A are distinct, then for the proper basis $\{\boldsymbol{x}_j\}$ of matrix A it is always possible to find a proper basis $\{\boldsymbol{x}_j'\}$ of the transpose A' such that

$$(\boldsymbol{x}_j, \boldsymbol{x}_k') = \delta_{jk} \tag{8}$$

where δ_{jk} is the Kronecker delta.

Corollary. If the matrix A is real and symmetric ($A' = A$), then we can put $\boldsymbol{x}'_j = \boldsymbol{x}_j$ ($j = 1, 2, \ldots, n$) where $\boldsymbol{x}_j$ are the normalized eigenvectors of A (see Sec. 10.15). Then

$$(\boldsymbol{x}_j, \boldsymbol{x}_k) = \delta_{jk}$$

Let us now derive the so-called *bilinear expansion of matrix* A.

Theorem 2. *Let* A *be a real square matrix and let*

$$X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

($j = 1, 2, \ldots, n$) *be its eigenvectors regarded as column matrices and let*

$$X'_k = [x'_{1k} \ \ldots \ x'_{nk}]$$

($k = 1, 2, \ldots, n$) *be the corresponding*[1] *eigenvectors of the transpose* A′ *regarded as row matrices; also let the conditions of biorthogonality* (8) *be met:*

$$(X_j, X_k'^{*}) = X'_k X_j = \delta_{jk} \tag{9}$$

Then the following relation is valid:

$$A = \lambda_1 X_1 X'_1 + \lambda_2 X_2 X'_2 + \ldots + \lambda_n X_n X'_n \tag{10}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ *are the eigenvalues of matrix* A.

Proof. We consider the matrices

$$X = \begin{bmatrix} x_{11} & \cdot & x_{1n} \\ \vdots & \cdot & \vdots \\ x_{n1} & \cdot & x_{nn} \end{bmatrix} \quad \text{and} \quad X' = \begin{bmatrix} x'_{11} & \ldots & x'_{n1} \\ \cdot & \cdot & \cdot \\ x'_{1n} & \ldots & x'_{nn} \end{bmatrix}$$

which consist of the columns X_j ($j = 1, \ldots, n$) and rows X'_k ($k = 1, \ldots, n$), respectively.

By (9) we have

$$X'X = \left[\sum_{k=1}^{n} x'_{ki} x_{kj}\right] = [X_j X'_i] = [\delta_{ji}] = E \tag{11}$$

where E is the unit matrix. Since the matrix X consists of linearly independent columns, it is nonsingular, that is, det $X \neq 0$ and, hence, there is an inverse X^{-1}. On the basis of (11) (see Sec. 7.4, Theorem, Note 1) we have

$$X^{-1} = X'$$

---

[1] That is, corresponding to the same eigenvalues of the matrices A and A'.

which implies that

$$XX' = E$$

and, thus, we obtain *the second set of biorthogonality relations* [7]:

$$\sum_{k=1}^{n} x_{ik}x'_{jk} = \delta_{ij} \tag{12}$$

Using these relations, we have

$$X_1X'_1 + X_2X'_2 + \ldots + X_nX'_n = [x_{i1}x'_{j1}] + [x_{i2}x'_{j2}] + \ldots + [x_{in}x'_{jn}] =$$

$$= \left[\sum_{k=1}^{n} x_{ik}x'_{jk}\right] = [\delta_{ij}] = E$$

That is,

$$E = X_1X'_1 + X_2X'_2 + \ldots + X_nX'_n$$

Premultiplying this equation by A and taking into account that

$$AX_j = \lambda_j X_j \quad (j = 1,\ 2,\ \ldots,\ n)$$

we will clearly get equation (10).

It will be well to note that in formula (10), X_j and X'_j $(j = 1, 2, \ldots, n)$ are the eigenvectors of the matrices A and A' corresponding to **one and the same** eigenvalue λ_j despite the notations of Theorem 1, where $\boldsymbol{x}_j$ and $\boldsymbol{x}'_j$ are the eigenvectors of the matrices A and A', which eigenvectors correspond to the **complex conjugate** eigenvalues λ_j and λ_j^*.

REFERENCES FOR CHAPTER 10

[1] *G. E. Shilov*, *Introduction to the Theory of Linear Spaces*, 1952, Chapters I-IX (in Russian).

[2] *I. M. Gelfand*, *Lectures on Linear Algebra*, 1951, Chapters I-II (in Russian).

[3] *A. I. Maltsev*, *Principles of Linear Algebra*, 1948, Chapters I-III (in Russian).

[4] *Alston S. Householder*, *Principles of Numerical Analysis*, 1953, Chapter 2.

[5] *Yu. A. Shreider*, *The Solution of Systems of Linear Algebraic Equations*, 1951 (in Russian).

[6] *F. R. Gantmacher*, *The Theory of Matrices*, 1953, Chapter VIII (in Russian); there is an English translation published by Chelsea, New York, 1959.

[7] *V. N. Faddeyeva*, *Computational Methods of Linear Algebra*, 1950, Chapter I (in Russian).

*Chapter 11

ADDITIONAL FACTS ABOUT THE CONVERGENCE OF ITERATION PROCESSES FOR SYSTEMS OF LINEAR EQUATIONS

11.1 THE CONVERGENCE OF MATRIX POWER SERIES

Theorem 1. *The matrix power series*

$$\sum_{k=0}^{\infty} a_k X^k \tag{1}$$

with numerical coefficients a_k *converges if all eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *of matrix* X *are located in the closed circle of convergence* $|x| \leqslant R$ (Fig. 56) *of the scalar power series*

$$\sum_{k=0}^{\infty} a_k x^k \tag{2}$$

$(x = \xi + i\eta)$; *and the eigenvalues on the circumference of the circle of convergence are simple and are the points of convergence of the series* (2).

Series (1) *diverges if even one eigenvalue of matrix* X *lies outside the closed circle of convergence of series* (2) *or if there is an eigenvalue of matrix* X *lying on the circumference of the circle of convergence for which series* (2) *diverges.*

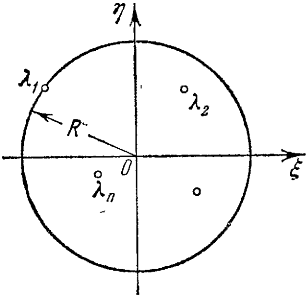


Fig. 56

Proof. (1) Let the matrix X be such that

$$|\lambda_1| \leqslant R, \ldots, |\lambda_n| \leqslant R$$

For the sake of simplicity we assume that the eigenvalues λ_j $(j = 1, 2, \ldots, n)$ of matrix X are simple. Then the matrix X can be reduced to diagonal form with the aid of the nonsingular matrix S:

$$X = S^{-1} [\lambda_1, \ldots, \lambda_n] S$$

Introducing the notation

$$F_m(X)=\sum_{k=0}^{m} a_k X^k, \qquad f_m(x)=\sum_{k=0}^{m} a_k x^k$$

and

$$f(x)=\lim_{m\to\infty} f_m(x)=\sum_{k=0}^{\infty} a_k x^k$$

we have

$$F_m(X)=\sum_{k=0}^{m} a_k\{S^{-1}[\lambda_1, \ldots, \lambda_n]S\}^k=$$

$$=S^{-1}\left\{\sum_{k=0}^{m} a_k[\lambda_1^k, \ldots, \lambda_n^k]\right\}S=S^{-1}[f_m(\lambda_1), \ldots, f_n(\lambda_n)]S \quad (3)$$

Since the eigenvalues λ_j lie within the circle of convergence of the power series (2) or coincide with the points of convergence of that series belonging to the circumference of the circle of convergence, there exist finite limits

$$f(\lambda_j)=\lim_{m\to\infty} f_m(\lambda_j) \qquad (j=1, 2, \ldots, n)$$

For this reason, passing to the limit in (3) as $m\to\infty$, we obtain

$$F(X)=\lim_{m\to\infty} F_m(X)=S^{-1}[f(\lambda_1), \ldots, f(\lambda_n)]S \quad (4)$$

Thus, the matrix series (1) converges at the point X.

It can likewise be proved that the assertion of the theorem is also true for the case of multiple eigenvalues λ_j, but we will not examine this case [1].

(2) Suppose, for instance, that at least one eigenvalue λ_1 of matrix X is such that

$$|\lambda_1|>R$$

Since λ_1 lies outside the circle of convergence of the power series (2), it follows that $f_m(\lambda_1)$ has no limit as $m\to\infty$. Formula (3) implies that $F_m(X)$ likewise has no limit as $m\to\infty$; thus, series (1) diverges at the point X.

A similar result is obtained if $|\lambda_1|=R$ and the series $\sum_{k=0}^{\infty} a_k\lambda_1^k$ diverges.

Note. Formula (4) implies that if $\lambda_1, \lambda_2, \ldots, \lambda_n$ are simple eigenvalues of matrix X, then $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)$, where

$$f(x)=\sum_{k=0}^{\infty} a_k x^k,$$

are eigenvalues of the function

$$F(X)=\sum_{k=0}^{\infty} a_k X^k$$

In particular, the numbers $\lambda_1^k, \ldots, \lambda_n^k$ are eigenvalues of the matrix X^k.

Theorem 2. *The geometric series of matrices*

$$E+X+X^2+\ldots+X^k+\ldots, \tag{5}$$

where X *is a square matrix of order* n, *converges if and only if all the eigenvalues*

$$\lambda_j=\lambda_j(X) \qquad (j=1, 2, \ldots, n)$$

of X *are located within the unit circle*

$$|\lambda_j|<1 \qquad (j=1, 2, \ldots, n) \tag{6}$$

If the series (5) *diverges, then* $X^k \not\to 0$ *as* $k\to\infty$.

Proof. Indeed, since for the appropriate power series

$$\sum_{k=0}^{\infty} x^k \tag{7}$$

the radius of convergence $R=1$, and for $|x|=1$ (7) diverges, then by Theorem 1 the geometric series (5) converges only when Conditions (6) hold true.

If the series (5) diverges, then

$$|\lambda_j|\geqslant 1 \qquad (j=1, 2, \ldots, n)$$

whence, assuming for the sake of simplicity that the eigenvalues $\lambda_1, \ldots, \lambda_n$ are distinct, we have

$$X=S^{-1}[\lambda_1, \ldots, \lambda_n]\,S$$

where S is a nonsingular matrix. Therefore

$$X^k=S^{-1}[\lambda_1^k, \ldots, \lambda_n^k]\,S$$

and so $X^k \not\to 0$ as $k\to\infty$. This assertion also holds true for multiple eigenvalues λ_j (we omit this part of the proof).

Theorem 3. *The modulus of each eigenvalue* $\lambda_1, \ldots, \lambda_n$ *of the square matrix* X *does not exceed any canonical norm of it:*

$$|\lambda_j|\leqslant\|X\| \qquad (j=1, 2, \ldots, n)$$

Proof. Put

$$\|X\|=\rho$$

and consider the matrix

$$Y = \frac{1}{\rho+\varepsilon} X \tag{8}$$

where $\varepsilon > 0$. Obviously

$$\|Y\| = \frac{1}{\rho+\varepsilon}\|X\| = \frac{\rho}{\rho+\varepsilon} < 1$$

Hence (see Sec. 7.10, Theorem 5) the series

$$E + Y + Y^2 + \dots + Y^k + \dots$$

converges.

From this we conclude, by Theorem 2, that the eigenvalues $\mu_1, \dots, \mu_n$ of matrix Y satisfy the inequalities

$$|\mu_j| < 1 \qquad (j = 1, 2, \dots, n)$$

But from formula (8) it follows that

$$\mu_j = \frac{1}{\rho+\varepsilon}\lambda_j \qquad (j = 1, 2, \dots, n)$$

Consequently

$$|\lambda_j| < \rho+\varepsilon \qquad (j = 1, 2, \dots, n)$$

or, because the number ε is arbitrary,

$$|\lambda_j| \leqslant \rho = \|X\| \qquad (j = 1, 2, \dots, n)$$

which completes the proof.

11.2 THE CAYLEY-HAMILTON THEOREM

Theorem. *Every square matrix* X *is a root of its characteristic polynomial; thus, if*

$$\psi(\lambda) = \lambda^n + p_1\lambda^{n-1} + \dots + p_n$$

where $\psi(\lambda) = \det(\lambda E - X)$, *then*

$$\psi(X) = X^n + p_1 X^{n-1} + \dots + p_n E \equiv 0$$

Proof. Let all the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of matrix X, that is the roots of the characteristic equation $\psi(\lambda) = 0$, be distinct. Then X may be reduced to diagonal form with the aid of some nonsingular matrix S:

$$X = S^{-1}[\lambda_1, \lambda_2, \dots, \lambda_n]S$$

Since $\psi(X)$ is a particular case of a matrix power series, we have, by formula (4) of Sec. 11.1,

$$\psi(X) = S^{-1}[\psi(\lambda_1), \psi(\lambda_2), \dots, \psi(\lambda_n)]S$$

But, clearly,

$$\psi(\lambda_j)=0 \qquad (j=1,\ 2,\ \ldots,\ n)$$

and so

$$\psi(X)=S^{-1}[0,\ 0,\ \ldots,\ 0]\,S=0$$

If the characteristic equation $\psi(\lambda)=0$ has multiple roots, then they may be regarded as the limits of noncoincident roots under infinitesimal perturbations of the coefficients of the equation [1]. As a result, the theorem can be generalized to this case as well.

11.3 NECESSARY AND SUFFICIENT CONDITIONS FOR THE CONVERGENCE OF THE PROCESS OF ITERATION FOR A SYSTEM OF LINEAR EQUATIONS

Using the eigenvalues of the matrix $\alpha=[\alpha_{ij}]$, it is possible to specify necessary and sufficient conditions for the convergence of the iteration process for a linear system

$$\boldsymbol{x}=\alpha\boldsymbol{x}+\boldsymbol{\beta} \tag{1}$$

where

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_1\\ \vdots\\ \beta_n\end{bmatrix} \quad\text{and}\quad \boldsymbol{x}=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$$

Theorem. *For the convergence of the iteration process*

$$\boldsymbol{x}^{(k)}=\alpha\boldsymbol{x}^{(k-1)}+\boldsymbol{\beta} \qquad (k=1,\ 2,\ \ldots) \tag{2}$$

for any choice of the initial vector $\boldsymbol{x}^{(0)}$ *and for any constant term* $\boldsymbol{\beta}$, *it is necessary and sufficient that the eigenvalues of the matrix* α, *that is the roots of the characteristic equation*

$$\det(\alpha-\lambda E)=0$$

be less than unity in modulus.

Proof. From formula (2) we get

$$\boldsymbol{x}^{(k)}=(E+\alpha+\alpha^2+\ldots+\alpha^{k-1})\,\boldsymbol{\beta}+\alpha^k\boldsymbol{x}^{(0)} \tag{3}$$

whence it follows that the convergence of the iteration process (2) for arbitrary $\boldsymbol{\beta}$ and $\boldsymbol{x}^{(0)}$ is equivalent to the convergence of the geometric series of matrices

$$E+\alpha+\alpha^2+\ldots=\sum_{k=0}^{\infty}\alpha^k \tag{4}$$

By Theorem 2 of Sec. 11.1, the geometric series (4) converges if all eigenvalues λ_j $(j=1,\ 2,\ \ldots,\ n)$ of the matrix α satisfy the inequalities

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n) \tag{5}$$

Since, in that case, $\alpha^k \to 0$ as $k \to \infty$, from formula (3) it follows that the process of iteration converges for arbitrary $\boldsymbol{\beta}$ and $\boldsymbol{x}^{(0)}$, that is there exists a limit

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}$$

where $\boldsymbol{x}$ is clearly a solution of system (1).

If inequalities (5) are not valid, then the series (4) diverges. In that case the iteration process will obviously diverge too for a certain choice of the initial vector $\boldsymbol{x}^{(0)}$.

Thus, for convergence of the process of iteration (2), it is necessary and sufficient that all the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the characteristic equation

$$\begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22}-\lambda & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn}-\lambda \end{vmatrix} = 0$$

satisfy the conditions: $|\lambda_j| < 1$ $(j = 1, 2, \ldots, n)$.

Corollary. For convergence of the process of iteration (2) it is sufficient that

$$\|\alpha\| < 1 \tag{6}$$

no matter what the canonical norm (cf. Sec. 9.1).

Indeed, we obtain inequalities (5) by virtue of Theorem 3 of Sec. 11.1 and on the basis of inequalities (6).

Note. Consider the linear system

$$A\boldsymbol{x} = \boldsymbol{b} \tag{7}$$

where $A = [a_{ij}]$ and $\boldsymbol{b} = [b_1 \ldots b_n]$ is a column vector.

Suppose

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix} \neq 0$$

To reduce (7) to the special form (1) we ordinarily set

$$A = D + (A - D)$$

whence

$$D\boldsymbol{x} = \boldsymbol{b} - (A - D)\,\boldsymbol{x}$$

and since $\det D = a_{11}a_{22}\ \dots\ a_{nn} \neq 0$, then

$$x = D^{-1}b + D^{-1}(D-A)x$$

We can take

$$\alpha = D^{-1}(D-A)$$

Thus, for the convergence of an ordinary iteration process for the linear system (7) given any constant term b and any initial vector $x^{(0)}$, it is necessary and sufficient that all the roots $\lambda_1, \lambda_2, \dots, \lambda_n$ of the characteristic equation

$$\det[D^{-1}(D-A) - \lambda E] = 0 \tag{8}$$

be less than unity in modulus. Taking advantage of the theorem on the determinant of a product of two matrices, equation (8) may be transformed as follows:

$$\det D^{-1} \det[(D-A) - \lambda D] = 0$$

or

$$\det[\lambda D + (A-D)] = 0$$

That is,

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}\lambda & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn}\lambda \end{vmatrix} = 0$$

11.4 NECESSARY AND SUFFICIENT CONDITIONS FOR THE CONVERGENCE OF THE SEIDEL PROCESS FOR A SYSTEM OF LINEAR EQUATIONS

Given the linear system

$$x = \alpha x + \beta \tag{1}$$

where $\alpha = [\alpha_{ij}]$ and $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$, consider the Seidel process

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k-1)} + \beta_i \quad (i = 1, 2, \dots, n;\quad k = 1, 2, \dots)$$

for an arbitrary initial vector

$$x^{(0)} = \begin{bmatrix} x_1^0 \\ \vdots \\ x_n^{(0)} \end{bmatrix}$$

Set

$$\alpha = B + C$$

where

$$B=\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ \alpha_{21} & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{n,\,n-1} & 0 \end{bmatrix}, \qquad C=\begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ 0 & \alpha_{22} & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \alpha_{nn} \end{bmatrix}$$

Then the Seidel process can be written in matrix form as follows:

$$x^{(k)} = Bx^{(k)} + Cx^{(k-1)} + \beta \qquad (k=1,\ 2,\ \dots) \tag{2}$$

Theorem. *For the convergence of the Seidel process* (2) *for system* (1) *given an arbitrary choice of constant term* β *and initial vector* $\mathbf{x}^{(0)}$, *it is necessary and sufficient that all the roots* $\lambda_1, \dots, \lambda_n$ *of the equation*

$$\det\,[C-(E-B)\lambda] \equiv \begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21}\lambda & \alpha_{22}-\lambda & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1}\lambda & \alpha_{n2}\lambda & \dots & \alpha_{nn}-\lambda \end{vmatrix} = 0 \tag{3}$$

be less than unity in modulus.

Proof. From formula (2) it follows that

$$(E-B)\,x^{(k)} = Cx^{(k-1)} + \beta \tag{4}$$

The matrix $E-B$ is nonsingular, since

$$\det\,(E-B) = 1$$

and so (4) can be written as

$$x^{(k)} = (E-B)^{-1}\,Cx^{(k-1)} + (E-B)^{-1}\,\beta \tag{5}$$

It is then clear that the Seidel process is equivalent to the process of simple iteration as applied to the linear system

$$x = (E-B)^{-1}\,Cx + (E-B)^{-1}\,\beta$$

By virtue of the theorem of the preceding section, for convergence of the iteration process (5) it is necessary and sufficient that the roots $\lambda_1, \dots, \lambda_n$ of the characteristic equation

$$\det\,[(E-B)^{-1}\,C-\lambda E] = 0 \tag{6}$$

satisfy the conditions

$$|\lambda_j| < 1 \qquad (j=1,\ 2,\ \dots,\ n)$$

Equation (6) is plainly equivalent to equation (3).

Note. Let

$$Ax = b \tag{7}$$

Set

$$D=\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \neq 0$$

In order to apply the Seidel method to (7), we ordinarily write it in the form

$$D\boldsymbol{x}=(D-A)\,\boldsymbol{x}+\boldsymbol{b}$$

or

$$\boldsymbol{x}=D^{-1}(D-A)\,\boldsymbol{x}+D^{-1}\boldsymbol{b} \tag{8}$$

Set

$$A-D=B_1+C_1$$

where

$$B_1=\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{n,\,n-1} & 0 \end{bmatrix}$$

and

$$C_1=\begin{bmatrix} 0 & a_{12} & \dots & a_{1,\,n-1} & a_{1n} \\ 0 & 0 & \dots & a_{2,\,n-1} & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & 0 & 0 \end{bmatrix}$$

Then

$$D^{-1}(D-A)=B+C$$

where

$$B=-D^{-1}B_1 \quad \text{and} \quad C=-D^{-1}C_1$$

and the triangular matrices B and C effect the partition of the matrix of system (8) that is necessary for application of the Seidel process. By formula (3), the convergence of the Seidel process for system (7) depends on the properties of the roots of the equation

$$\det\,[-D^{-1}C_1-(E+D^{-1}B_1)\,\lambda]=0 \tag{9}$$

Equation (9) can be replaced by the equivalent equation

$$\det\,[(D+B_1)\,\lambda+C_1]=0$$

or

$$\begin{vmatrix} a_{11}\lambda & a_{12}\lambda & \dots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}\lambda & a_{n2}\lambda & \dots & a_{nn} \end{vmatrix}=0 \tag{10}$$

Thus, for the convergence of the Seidel process for system (7) given an arbitrary constant term $\boldsymbol{b}$ and an arbitrary initial approximation $\boldsymbol{x}^{(0)}$, it is necessary and sufficient that all the roots λ_j of equation (10) satisfy the conditions

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n)$$

11.5 CONVERGENCE OF THE SEIDEL PROCESS FOR A NORMAL SYSTEM

Theorem. *For a normal system, the ordinary Seidel process converges for any choice of the initial vector.*

Proof. Let the linear system

$$A\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

be normal, that is, let the matrix $A = [a_{ij}]$ be symmetric and positive definite.

Set

$$A = D + V + V^*$$

where

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix}$$

is a diagonal matrix,

$$V = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & 0 \end{bmatrix}$$

is a lower triangular matrix, and

$$V^* = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

is an upper triangular matrix, which, because A is symmetric, is the transpose of V. We then have

$$(D + V + V^*)\,\boldsymbol{x} = \boldsymbol{b}$$

whence

$$D\boldsymbol{x} = \boldsymbol{b} - (V + V^*)\,\boldsymbol{x}$$

and, consequently,

$$x = D^{-1}b - D^{-1}(V+V^*)x \tag{2}$$

where

$$D^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \frac{1}{a_{nn}} \end{bmatrix}$$

According to the foregoing, the Seidel process for system (1) or the equivalent system (2) is constructed as follows:

$$x^{(k)} = D^{-1}b + Bx^{(k)} + Cx^{(k-1)} \qquad (k = 1, 2, \dots) \tag{3}$$

where

$$B = -D^{-1}V \text{ and } C = -D^{-1}V^*$$

By virtue of the theorem of the preceding section, for the process to converge it is necessary and sufficient that all the eigenvalues λ of the matrix

$$M = (E-B)^{-1}C$$

be less than unity in modulus.

In our case we have

$$M = -(E+D^{-1}V)^{-1}D^{-1}V^* = -[D^{-1}(D+V)]^{-1}D^{-1}V^* =$$
$$= -(D+V)^{-1}DD^{-1}V^* = -(D+V)^{-1}V^*$$

Let e be a unit eigenvector of matrix M corresponding to the eigenvalue λ, that is,

$$(D+V)^{-1}V^*e = -\lambda e$$

or

$$V^*e = -\lambda(D+V)e$$

From this,

$$(V^*e, e) = -\lambda[(D+V)e, e]$$

and hence

$$\lambda = -\frac{(V^*e, e)}{(De, e)+(Ve, e)}$$

We introduce the notation

$$(De, e) = \sum_{j=1}^{n} a_{jj}|e_j|^2 = \sigma > 0$$

and

$$(Ve, e) = \alpha + i\beta$$

where α and β are real and $i^2 = -1$.

Since the matrix V^* is the transpose of V, we obtain

$$(V^*e,\ e)=(e,\ Ve)=(Ve,\ e)^*=\alpha-i\beta$$

Therefore

$$\lambda=-\frac{\alpha-i\beta}{(\sigma+\alpha)+i\beta}$$

and hence

$$|\lambda|=\frac{\sqrt{\alpha^2+\beta^2}}{\sqrt{(\sigma+\alpha)^2+\beta^2}} \tag{4}$$

Using the positive definiteness of the matrix A, we get

$$(Ae,\ e)=(De,\ e)+(Ve,\ e)+(V^*e,\ e)=$$
$$=\sigma+(\alpha+i\beta)+(\alpha-i\beta)=\sigma+2\alpha>0$$

that is,

$$\sigma+\alpha>-\alpha \tag{5}$$

Furthermore, taking into account the positive nature of σ, we clearly have

$$\sigma+\alpha>\alpha$$

Thus, the inequality

$$\sigma+\alpha>|\alpha| \tag{6}$$

is always valid, whence for the terms of the fraction (4) we get

$$\sqrt{(\sigma+\alpha)^2+\beta^2}>\sqrt{\alpha^2+\beta^2}\geqslant 0$$

or

$$|\lambda|<1$$

which is what we set out to prove.

11.6 METHODS FOR EFFECTIVELY CHECKING THE CONDITIONS OF CONVERGENCE

In order to verify the conditions of the theorems of convergence of iteration processes, it is necessary to have effective criteria that permit determining whether the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of a given algebraic polynomial

$$f(\lambda)=p_0\lambda^n+p_1\lambda^{n-2}+\ldots+p_n \tag{1}$$

meet the requirement

$$|\lambda_j|<1 \qquad (j=1,\ 2,\ \ldots,\ n) \tag{2}$$

or do not. This problem is settled very simply by using the well-known *Hurwitz conditions* [2].

Theorem (Hurwitz). *Let the coefficients* p_k $(k=0, 1, \ldots, n)$ *of the polynomial* (1) *be real, and*

$$p_0 > 0$$

and let

$$M=\left[\begin{array}{cc|cccccc|c}
p_1 & p_0 & 0 & 0 & \ldots & 0 & 0 & & 0 \\
p_3 & p_2 & p_1 & p_0 & \ldots & 0 & 0 & & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot\ \cdot \\
0 & 0 & 0 & 0 & \ldots & p_n & p_{n-1} & & p_{n-2} \\
\hline
0 & 0 & 0 & 0 & \ldots & 0 & 0 & & p_n
\end{array}\right]$$

be a matrix of order n *whose rows are extended sequences of the coefficients of polynomial* (1),

$$p_{2m-1}, \ p_{2m-2}, \ \ldots, \ p_{2m-n}$$

where $p_k=0$ *for* $k<0$ *and* $k>n$. *Then all roots* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *of* (1) *have negative real parts*

$$\operatorname{Re}\lambda_j < 0 \qquad (j=1, 2, \ldots, n)$$

(which is to say they are located in the left half-plane of the complex plane $\lambda=\alpha+i\beta$*) if and only if the principal diagonal minors of matrix* M *are positive; that is*

$$\left.\begin{array}{l}
\Delta_1 = p_1 > 0 \\
\Delta_2 = \begin{vmatrix} p_1 p_0 \\ p_3 p_2 \end{vmatrix} > 0, \\
\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\
\Delta_n = p_n \Delta_{n-1} > 0
\end{array}\right\} \tag{3}$$

Example 1. For the quadratic trinomial

$$p_0\lambda^2 + p_1\lambda + p_2$$

the Hurwitz conditions are

$$p_0 > 0, \quad p_1 > 0, \quad p_2 > 0$$

We would like to know when the roots of polynomial (1) satisfy Condition (2), that is, when they lie in the complex plane λ within the unit circle

$$|\lambda| < 1$$

Using the linear-fractional function

$$\lambda = \frac{\mu+1}{\mu-1}$$

we can transform the interior of the unit circle $|\lambda|<1$ into the left half-plane $\operatorname{Re}\mu < 0$. Our polynomial (1) then becomes

$$f\left(\frac{\mu+1}{\mu-1}\right) = p_0\left(\frac{\mu+1}{\mu-1}\right)^n + p_1\left(\frac{\mu+1}{\mu-1}\right)^{n-1} + \ldots + p_n =$$
$$= \frac{1}{(\mu-1)^n}\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots + p_n(\mu-1)^n\right]$$

Hence, the roots of (1) are located inside the unit circle if and only if the auxiliary polynomial

$$F(\mu) = \pm\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots + p_n(\mu-1)^n\right]$$

satisfies the Hurwitz conditions (3), and the sign is chosen so that the leading coefficient

$$\pm(p_0 + p_1 + \ldots + p_n) > 0$$

Example 2. Consider the quadratic trinomial

$$f(\lambda) = \lambda^2 + p\lambda + q \tag{4}$$

where p and q are real. The auxiliary polynomial is of the form

$$F(\mu) = \pm\left[(\mu+1)^2 + p(\mu+1)(\mu-1) + q(\mu-1)^2\right] =$$
$$= \pm\left[(1+p+q)\mu^2 + 2(1-q)\mu + (1-p+q)\right]$$

From the Hurwitz conditions we get

$$\left.\begin{aligned} &\pm(1+p+q) > 0, \\ &\pm(1-q) > 0, \\ &\pm(1-p+q) > 0 \end{aligned}\right\}$$

Consider the cases:

(a) $q < 1$, then $q > -p-1$ and $q > p-1$;

(b) $q > 1$, then $q < -p-1$ and $q < p-1$, which is impossible.

Consequently, equation (4) has the roots $\lambda_1\lambda_2$ which are less than unity in modulus if and only if

$$|p| < 1+q, \quad |q| < 1 \tag{5}$$

Since for $n=2$ the characteristic equation of the matrix α is of the form

$$\begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} \\ \alpha_{21} & \alpha_{22}-\lambda \end{vmatrix} = 0$$

or

$$\lambda^2 - (\alpha_{11}+\alpha_{22})\lambda + \det\alpha = 0$$

then for the convergence of the appropriate iteration process for a system of two equations it is necessary that

$$|\det\alpha| < 1$$

The regions of convergence of the process of ordinary iteration and the process of Seidel overlap, generally speaking. There are cases of linear systems for which the ordinary iteration process converges while the Seidel process diverges, and conversely [3].

Example 3. Consider the linear system

$$x = \alpha x + \beta \tag{6}$$

with the matrix

$$\alpha = \begin{bmatrix} p & q \\ -q & p \end{bmatrix}$$

where p and q are real.

The characteristic equation of matrix α is of the form

$$\begin{vmatrix} p-\lambda & q \\ -q & p-\lambda \end{vmatrix} = 0$$

or

$$(\lambda - p)^2 + q^2 = 0$$

whence

$$\lambda_{1,2} = p \pm iq$$

For the convergence of the process of ordinary iteration it is necessary and sufficient that

$$|\lambda_{1,2}| = \sqrt{p^2+q^2} < 1$$

that is

$$p^2 + q^2 < 1$$

(region A in Fig. 57).

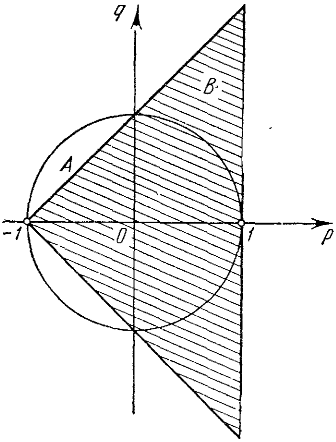


Fig. 57

For the Seidel method, the equation defining convergence is of the form

$$\begin{vmatrix} p-\lambda & q \\ -q\lambda & p-\lambda \end{vmatrix} = 0$$

or

$$\lambda^2 - (2p - q^2)\lambda + p^2 = 0 \tag{7}$$

On the basis of the results of Example 2, for the roots λ_1 and λ_2 of equation (7) to satisfy the conditions

$$|\lambda_1| < 1, \qquad |\lambda_2| < 1$$

it is necessary and sufficient that the following inequalities be valid:

$$|2p-q^2| < 1+p^2, \qquad p^2 < 1$$

whence

$$|p| < 1, \qquad |q| < 1+p$$

(region B in Fig. 57). Since the regions A and B partially overlap, it follows that for system (6) one can choose the coefficients p and q, firstly, so that the process of iteration converges and the Seidel process diverges (for example, $p=-0.5$, $q=0.6$), and, secondly, so that, conversely, the Seidel process converges and the iteration process diverges (for example, $p=0.5$, $q=1$).

REFERENCES FOR CHAPTER 11

[1] *V. I. Smirnov, Course of Higher Mathematics*, 1933, Chapter VII (in Russian).

[2] *A. G. Kurosh, Course of Higher Algebra*, 1972, Chapter 8 (translated from the Russian).

[3] *V. N Faddeyeva, Computational Methods of Linear Algebra*, 1950, Chapter II, (in Russian).

Chapter 12

FINDING THE EIGENVALUES AND EIGENVECTORS OF A MATRIX

12.1 INTRODUCTORY REMARKS

In solving theoretical and practical problems, one often finds it necessary to determine the eigenvalues of a given matrix A, that is, to compute the roots of its *characteristic (secular) equation*

$$\det(A-\lambda E)=0 \tag{1}$$

and also to find the corresponding eigenvectors of A. The second problem is simpler because if the roots of the characteristic equation are known, then finding the eigenvectors reduces to finding nontrivial solutions of certain homogeneous linear systems. We therefore begin with the first problem: to compute the roots of the characteristic equation (1).

Here, two techniques are chiefly used: (1) expanding the secular determinant into a polynomial of degree n:

$$D(\lambda)=\det(A-\lambda E)$$

and then solving the equation $D(\lambda)=0$ by one of the familiar approximate methods (like, say, the Lobachevsky-Graeffe method described in Secs. 5.7 to 5.12) and (2) approximating the roots of the characteristic equation (mostly the numerically largest ones) by the method of iteration without any preliminary expansion of the secular determinant.

In this chapter we discuss the principal methods of solving the given general problem and begin with the *expansion of secular determinants*.

12.2 EXPANSION OF SECULAR DETERMINANTS

As is well known, *the secular determinant* of a matrix $A=[a_{ij}]$ is a determinant of the form

$$D(\lambda)=\det(A-\lambda E)=\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix} \tag{1}$$

Equating this determinant to zero we get the *characteristic equation*

$$D(\lambda)=0$$

If it is required to find all the roots of the characteristic equation, then it is desirable, first of all, to compute determinant (1).

Expanding determinant (1), we get an nth-degree polynomial:

$$D(\lambda)=(-1)^n[\lambda^n-\sigma_1\lambda^{n-1}+\sigma_2\lambda^{n-2}-\ldots+(-1)^n\sigma_n] \qquad (2)$$

where

$$\sigma_1=\sum_{\alpha=1}^{n}a_{\alpha\alpha}$$

is the sum of all first-order diagonal minors of the matrix A,

$$\sigma_2=\sum_{\alpha<\beta}\begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}$$

is the sum of all second-order diagonal minors of A,

$$\sigma_3=\sum_{\alpha<\beta<\gamma}\begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma} \end{vmatrix}$$

is the sum of all third-order diagonal minors of matrix A, and so forth. Finally,

$$\sigma_n=\det A$$

It is easy to see that the number of kth-order diagonal minors of A is

$$C_n^k=\frac{n(n-1)\ldots(n-k+1)}{k!} \qquad (k=1,\ 2,\ \ldots,\ n)$$

From this we find that direct computation of the coefficients of the characteristic polynomial (2) is equivalent to computing

$$C_n^1+C_n^2+\ldots+C_n^n=2^n-1$$

determinants of various orders, which, generally speaking, is a problem that is hard to handle when the values of n are large. This has given rise to special methods for expanding secular determinants (the methods of A. N. Krylov, A. M. Danilevsky, Leverrier, the method of undetermined coefficients, the method of interpolation, and others) (see [1]). Some of these methods will be examined in the following sections.

12.3 THE METHOD OF DANILEVSKY

The essence of the Danilevsky method [1] consists in reducing the secular determinant to the so-called *Frobenius standard form:*

$$D(\lambda)=\begin{vmatrix} p_1-\lambda & p_2 & p_3 & \dots & p_n \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & -\lambda \end{vmatrix} \tag{1}$$

If we succeed in writing the secular determinant in the form (1), then, after expanding it in terms of the elements of the first row, we get

$$D(\lambda)=(p_1-\lambda)(-\lambda)^{n-1}-p_2(-\lambda)^{n-2}+p_3(-\lambda)^{n-3}-\dots+(-1)^{n-1}p_n$$

or

$$D(\lambda)=(-1)^n(\lambda^n-p_1\lambda^{n-1}-p_2\lambda^{n-2}-p_3\lambda^{n-3}-\dots-p_n) \tag{2}$$

Thus, expanding a secular determinant written in the normal form (1) does not present any difficulties. Denote the given matrix by

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

and the *Frobenius matrix* similar to it by

$$P=\begin{bmatrix} p_1 & p_2 & \dots & p_{n-1} & p_n \\ 1 & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

Thus

$$P=S^{-1}AS$$

where S is a nonsingular matrix.

Since similar matrices have the same characteristic polynomials, we have

$$\det(A-\lambda E)=\det(P-\lambda E) \tag{3}$$

Thus, to substantiate this method, it will suffice to show how matrix P is constructed by proceeding from matrix A. According to the Danilevsky method, the transition from matrix A to the similar matrix P is effected by means of $n-1$ similarity transformations which successively transform the rows of A, beginning with the last, into corresponding rows of P.

Let us illustrate the beginning of the process. Our purpose is to carry the row

$$a_{n1}a_{n2} \dots a_{n,n-1}a_{nn}$$

into the row 0 0 ... 1 0. Assuming that $a_{n,n-1} \neq 0$, divide all elements of the $(n-1)$th column of A by $a_{n,n-1}$. Then its nth row will take the form

$$a_{n1}a_{n2} \dots 1a_{nn}$$

Then subtract from all the other columns the $(n-1)$th column of the transformed matrix multiplied by the numbers $a_{n1}, a_{n2}, \dots, a_{nn}$, respectively.

We thus obtain a matrix whose last row is of the desired form: 0 0 ... 1 0. The foregoing operations are elementary transformations performed on the columns of matrix A. Performing these same transformations on the unit matrix, we get the matrix

$$M_{n-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m_{n-1,1} & m_{n-1,2} & \dots & m_{n-1,n-1} & m_{n-1,n} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

where

$$m_{n-1,i} = -\frac{a_{ni}}{a_{n,n-1}} \quad \text{for} \quad i \neq n-1 \tag{4}$$

and

$$m_{n-1,n-1} = \frac{1}{a_{n,n-1}} \tag{4'}$$

From this we conclude (see Sec. 7.14) that these operations are equivalent to postmultiplying matrix M_{n-1} by matrix A; that is, the foregoing transformations result in the matrix

$$AM_{n-1} = B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1,n-1} & b_{1,n} \\ b_{21} & b_{22} & \dots & b_{2,n-1} & b_{2,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{n-1,1} & b_{n-1,2} & \dots & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & & 1 & 0 \end{bmatrix} \tag{5}$$

Using the rule of matrix multiplication, we find that the elements of B are computed from the following formulas:

$$b_{ij} = a_{ij} + a_{i,n-1}m_{n-1,j} \quad \text{for} \quad 1 \leqslant i \leqslant n, \quad j \neq n-1, \tag{6}$$

$$b_{j,n-1} = a_{i,n-1}m_{n-1,n-1} \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{6'}$$

However the matrix $B = AM_{n-1}$ will not be similar to A. To have a similarity transformation, it is necessary to postmultiply the inverse matrix M_{n-1}^{-1} by the matrix B:

$$M_{n-1}^{-1}AM_{n-1} = M_{n-1}^{-1}B$$

It is readily seen by direct verification that the inverse matrix M_{n-1}^{-1} is of the form

$$M_{n-1}^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \tag{7}$$

Let

$$M_{n-1}^{-1}AM_{n-1} = C$$

Then

$$C = M_{n-1}^{-1}B \tag{8}$$

Since it is quite obvious that postmultiplication of matrix M_{n-1}^{-1} by matrix B does not alter the transformed row of B, matrix C is of the form

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1,n-1} & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2,n-1} & c_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ c_{n-1,1} & c_{n-1,2} & \dots & c_{n-1,n-1} & c_{n-1,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{9}$$

Multiplying matrices M_{n-1}^{-1} (7) and B (5) together, we get

$$c_{ij} = b_{ij} \quad \text{for} \quad 1 \leqslant i \leqslant n-2 \tag{10}$$

and

$$c_{n-1,j} = \sum_{k=1}^{n} a_{nk}b_{kj} \quad \text{for} \quad 1 \leqslant j \leqslant n \tag{10'}$$

Thus, multiplication of M_{n-1}^{-1} by B only changes the $(n-1)$th row of B. The elements of this row are found from formulas (10) and (10′). The resulting matrix C is similar to A and has one reduced row. This concludes the first stage of the process.

Now, if $c_{n-1,n-2} \neq 0$, then similar operations are performed on matrix C by taking its $(n-2)$th row as the principal one. We then obtain the matrix

$$D = M_{n-2}^{-1}CM_{n-2}$$

with two reduced rows. This matrix is subjected to the same operations, and the process is continued until we finally obtain the

Frobenius matrix

$$P = M_1^{-1} \dots M_{n-2}^{-1} M_{n-1}^{-1} A M_{n-1} M_{n-2} \dots M_1$$

if, of course, all the $n-1$ intermediate transformations are possible.

The entire process can be arranged in a convenient computational scheme, the formation of which is illustrated by the following example.

Example. Reduce the following matrix to the Frobenius form:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

Solution. The computations are arranged in Table 25.

Enter the elements a_{ij} $(i, j = 1, 2, 3, 4)$ of the given matrix and the check sums $a_{i5} = \sum_{j=1}^{n} a_{ij}$ $(i = 1, 2, 3, 4)$ $(\sum)$ in rows 1 to 4 of Table 25. We mark element $a_{43} = 2$ in the third column (*marked column*). In row I we enter the elements of the third row of the matrix $M_{n-1} = M_3$ computed from formulas (4) and (4′):

$$m_{31} = -\frac{a_{41}}{a_{43}} = -\frac{4}{2} = -2,$$

$$m_{32} = -\frac{a_{42}}{a_{43}} = -\frac{3}{2} = -1.5,$$

$$m_{33} = \frac{1}{a_{43}} = \frac{1}{2} = 0.5,$$

$$m_{34} = -\frac{a_{44}}{a_{43}} = -\frac{1}{2} = -0.5$$

Here also (Row I of Table 25) we enter the element

$$m_{35} = -\frac{a_{45}}{a_{43}} = -\frac{10}{2} = -5$$

which is obtained by a similar device from the check column $\sum$. The number -5 must coincide with the sum of the elements of Row I that do not enter the check column (after element m_{33} is replaced by -1). For the sake of convenience, we write the number -1 alongside the element m_{33} and separate it from the latter by a dash.

In rows 5 to 8 of column M^{-1} we enter the third row of the matrix M^{-1}, which, by virtue of formula (7), coincides with the fourth row of the original matrix A. In the appropriate columns

TABLE 25
COMPUTATIONAL SCHEME OF DANILEVSKY'S METHOD

Row number	M^{-1}	Columns of matrix				Σ	Σ
		1	2	3	4		
1		1	2	3	4	10	
2		2	1	2	3	8	
3		3	2	1	2	8	
4		4	3	$\boxed{2}$	1	10	
I	$\overline{M_3^{-1}}$ M_3	−2	−1.5	$\overline{0.5}$ −1	−0.5	−5	
5	4	−5	−2.5	1.5	2.5	−3.5	−5
6	3	2	−2	1	2	−1	−2
$\boxed{7}$	2	1	0.5	0.5	1.5	3.5	3
8	1	0	0	1	0	1	0
$\boxed{7'}$		−24	$\boxed{-15}$	11	19	−9	
II	$\overline{M_2^{-1}}$ M_2	−1.600	$\overline{-0.067}$ −1	0.733	1.267	−0.600	
9	−24	−1	0.167	−0.333	−0.667	−1.833	−2
$\boxed{10}$	−15	1.2	0.133	−0.467	−0.533	0.333	0.2
11	11	0	1	0	0	1	0
12	19	0	0	1	0	1	1
$\boxed{10'}$		$\boxed{6}$	5	34	24	69	
III	$\overline{M_1^{-1}}$ M_1	$\overline{0.167}$ −1	−0.833	−5.667	−4.000	−11.500	
$\boxed{13}$	6	−0.167	1	5.333	3.333	9.500	9.667
14	5	1	0	0	0	1	0
15	34	0	1	0	0	1	1
16	24	0	0	1	0	1	1
$\boxed{13'}$		4	40	56	20	120	

of rows 5 to 8 we enter the elements of the matrix

$$B = AM_3$$

which are computed from the two-term formulas (6) for nonmarked columns and by the one-term formula (6) for the marked column. Thus, for the first column we have

$$\begin{aligned} b_{11} &= 1+3(-2) = -5, \\ b_{21} &= 2+2(-2) = -2, \\ b_{31} &= 3+1(-2) = 1, \\ b_{41} &= 4+2(-2) = 0 \end{aligned}$$

and so forth.

The transformed elements of the third (marked) column are obtained by multiplying the initial elements by $m_{33} = 0.5$. For example,

$$\begin{aligned} b_{13} &= 3 \cdot 0.5 = 1.5, \\ b_{23} &= 2 \cdot 0.5 = 1, \\ b_{33} &= 1 \cdot 0.5 = 0.5, \\ b_{43} &= 2 \cdot 0.5 = 1 \end{aligned}$$

Note that the last row of matrix B must have the form

$$0 \quad 0 \quad 1 \quad 0$$

For a check, we augment the matrix B with transformations by analogous two-term formulas with $m_{35} = -5$ corresponding elements of the column marked $\sum$. For instance,

$$\begin{aligned} b_{16} &= 10+3 \cdot (-5) = -5, \\ b_{26} &= 8+2 \cdot (-5) = -2, \\ b_{36} &= 8+1 \cdot (-5) = 3, \\ b_{46} &= 10+2 \cdot (-5) = 0 \end{aligned}$$

The results obtained are entered in the $\sum$ column in appropriate rows. Adding to them the elements of the third column, we have the check sums

$$b_{i5} = \sum_{j=1}^{4} b_{ij} \qquad (i = 1, 2, 3, 4)$$

for the rows 5-8 (column $\sum$).

The transformation M_3^{-1} which was performed on the matrix B and which yielded the matrix $C = M_3^{-1}B$ only alters the third row of B, that is, the seventh row of the table. The elements of this transformed row 7′ are obtained from formula (10); they are thus the sums of the paired products of the elements of column M^{-1}

located in rows 5 to 8 by the corresponding elements of each of the columns of matrix B. For example,

$$c_{31} = 4(-5) + 3(-2) + 2 \cdot 1 = -24$$

and so on.

We perform the same transformations on the $\sum$ column:

$$c_{35} = 4(-3.5) + 3(-1) + 2 \cdot 3.5 + 1 \cdot 1 = -9$$

This way we get a matrix C consisting of rows 5, 6, 7′, 8 with check sums $\sum$, and C is similar to A and has one reduced row 8. This completes the construction of the first similarity transformation $C = M_3^{-1}AM_3$.

Now, taking matrix C as the initial matrix and isolating element $c_{32} = -15$ (second column), we continue the process as before. We thus get the matrix $D = M_2^{-1}CM_2$, the elements of which are located in rows 9, 10′, 11, 12, which matrix contains two reduced rows. Finally, starting from element $d_{21} = 6$ (first column) and transforming the matrix D into a similar matrix, we get the desired Frobenius matrix P, whose elements are entered in the rows 13′, 14, 15, and 16. Checking at each stage in the process is effected via columns $\sum$ and $\sum'$.

Thus, the Frobenius matrix is

$$P = \begin{bmatrix} 4 & 40 & 56 & 20 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

whence the secular determinant reduced to the Frobenius standard form will be written as

$$D(\lambda) = \begin{vmatrix} 4-\lambda & 40 & 56 & 20 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{vmatrix}$$

or

$$D(\lambda) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20$$

12.4 EXCEPTIONAL CASES IN THE DANILEVSKY METHOD

The Danilevsky process proceeds without any complications if the chosen elements are nonzero. We will now examine exceptional cases when this requirement is not met.

Suppose that in transforming matrix A into the Frobenius matrix P we arrived, after a few steps, at a matrix of the form

$$D=\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} & \dots & d_{1,n-1} & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2k} & \dots & d_{2,n-1} & d_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ d_{k1} & d_{k2} & \dots & d_{kk} & \dots & d_{k,n-1} & d_{kn} \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & \dots & 1 & 0 \end{bmatrix}$$

and it was found that $d_{k,k-1}=0$.

It is then impossible to continue the transformation by the Danilevsky method. Two cases are possible here.

1. Let some element of matrix D to the left of the zeroth element $d_{k,k-1}$ be different from zero, that is, $d_{k,l}\neq 0$, where $l<k-1$. We then move this element to the position of the zeroth element $d_{k,k-1}$; that is, we interchange the $(k-1)$th and lth columns of D and simultaneously interchange the $(k-1)$th and lth rows. It can be proved that the resulting new matrix D' will be similar to the earlier one. We then apply the Danilevsky method to this new matrix.

2. Suppose $d_{kl}=0$ $(l=1, 2, \dots, k-1)$, then D is of the form

$$D=\left[\begin{array}{cccc|cccc} & (D_1) & & & & (L) & & \\ c_{11} & c_{12} & \dots & c_{1,k-1} & c_{1k} & \dots & c_{1,n-1} & c_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{k-1,1} & c_{k-1,2} & \dots & c_{k-1,k-1} & c_{k-1,k} & \dots & c_{k-1,n-1} & c_{k-1,n} \\ \hline 0 & 0 & \dots & 0 & c_{kk} & \dots & c_{k,n-1} & c_{kn} \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 \\ & (0) & & & & (D_2) & & \end{array}\right]=$$

$$=\left[\begin{array}{c|c} D_1 & L \\ \hline 0 & D_2 \end{array}\right]$$

In this case the secular determinant $\det(D-\lambda E)$ breaks up into two determinants:

$$\det(D-\lambda E)=\det(D_1-\lambda E)\det(D_2-\lambda E)$$

Here, the matrix D_2 is already reduced to the canonical form of Frobenius and so $\det(D_2-\lambda E)$ is computed at once. It remains to apply the Danilevsky method to the matrix D_1.

12.5 COMPUTATION OF EIGENVECTORS BY THE DANILEVSKY METHOD

The method of Danilevsky permits finding the eigenvectors of a given matrix A if the eigenvalues are known. Let λ be an eigenvalue of matrix A and, hence, an eigenvalue of the Frobenius matrix P which is similar to it.

We find the eigenvector $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ of P corresponding to the given eigenvalue λ: $P\boldsymbol{y}=\lambda\boldsymbol{y}$, whence $(P-\lambda E)\boldsymbol{y}=\mathbf{0}$ or

$$\begin{bmatrix} p_1-\lambda & p_2 & p_3 & \ldots & p_n \\ 1 & -\lambda & 0 & \ldots & 0 \\ 0 & 1 & -\lambda & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & -\lambda \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = 0$$

Multiplying the matrices together, we get a system for determining the coordinates $y_1, y_2, \ldots, y_n$ of the eigenvector $\boldsymbol{y}$:

$$\left.\begin{aligned} (p_1-\lambda)y_1+p_2y_2+\ldots+p_ny_n &= 0, \\ y_1-\lambda y_2 \qquad &= 0, \\ y_2-\lambda y_3 \qquad &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ y_{n-1}-\lambda y_n &= 0 \end{aligned}\right\} \qquad (1)$$

System (1) is homogeneous. Its values, to within the proportionality factor, can be found as follows. Put $y_n=1$. We then successively obtain

$$\left.\begin{aligned} y_{n-1} &= \lambda, \\ y_{n-2} &= \lambda^2, \\ \ldots & \\ y_1 &= \lambda^{n-1} \end{aligned}\right\} \qquad (2)$$

Thus, the desired eigenvector is

$$\boldsymbol{v}=\begin{bmatrix} \lambda^{n-1} \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}$$

Now denote by $\boldsymbol{x}$ the eigenvector of matrix A corresponding to the eigenvalue λ. It is clear then that we have

$$\boldsymbol{x}=M_{n-1}M_{n-2}\ldots M_2M_1\boldsymbol{y}$$

The transformation M_1 performed on $\boldsymbol{y}$ yields

$$M_1\boldsymbol{y}=\begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} \sum\limits_{k=1}^{n} m_{1k}y_k \\ y_2 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} \sum\limits_{k=1}^{n} m_{1k}y_k \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}$$

Thus, the transformation M_1 only alters the first coordinate of the vector. Similarly, the transformation M_2 alters only the second coordinate of the vector $M_1\boldsymbol{y}$, etc. Repeating this process $n-1$ times, we obtain the desired eigenvector $\boldsymbol{x}$ of matrix A.

12.6 THE METHOD OF KRYLOV

We will now examine a method of expanding a secular determinant that is due to A. N. Krylov [2] and is based on an essentially different principle than the method of Danilevsky.

Let

$$D(\lambda)\equiv\det(\lambda E-A)=\lambda^n+p_1\lambda^{n-1}+\ldots+p_n \tag{1}$$

be the characteristic polynomial (apart from sign) of the matrix A. By the Cayley-Hamilton theorem (Sec. 11.2), matrix A reduces its characteristic polynomial to zero, and so

$$A^n+p_1A^{n-1}+\ldots+p_nE=0 \tag{2}$$

Now let us take an arbitrary nonzero vector

$$\boldsymbol{y}^{(0)}=\begin{bmatrix} y_1^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix}$$

Postmultiplying both members of (2) by $\boldsymbol{y}^{(0)}$, we get

$$A^n\boldsymbol{y}^{(0)}+p_1A^{n-1}\boldsymbol{y}^{(0)}+\ldots+p_n\boldsymbol{y}^{(0)}=\mathbf{0} \tag{3}$$

Set

$$A^k\boldsymbol{y}^{(0)}=\boldsymbol{y}^{(k)} \quad (k=1,\ 2,\ \ldots,\ n) \tag{4}$$

Then equation (3) takes the form

$$\mathbf{y}^{(n)}+p_1\mathbf{y}^{(n-1)}+\ldots+p_n\mathbf{y}^{(0)}=0 \tag{5}$$

or

$$\begin{bmatrix} y_1^{(n-1)} & y_1^{(n-2)} & \vdots & y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} & \vdots & y_2^{(0)} \\ \vdots & \vdots & & \vdots \\ y_n^{(n-1)} & y_n^{(n-2)} & \vdots & y_n^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = -\begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix} \tag{5'}$$

where

$$\mathbf{y}^{(k)}=\begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_n^{(k)} \end{bmatrix} \qquad (k=0,\ 1,\ 2,\ \ldots,\ n)$$

Hence, the vector equation (5) is equivalent to the system of equations

$$p_1y_j^{(n-1)}+p_2y_j^{(n-2)}+\ldots+p_ny_j^{(0)}=-y_j^{(n)} \qquad (j=1,\ 2,\ \ldots,\ n) \tag{6}$$

from which we can, generally speaking, determine the unknown coefficients $p_1,\ p_2,\ \ldots,\ p_n$.

Since on the basis of formula (4)

$$\mathbf{y}^{(k)}=A\mathbf{y}^{(k-1)}$$

$(k=1,\ 2,\ \ldots,\ n)$, the coordinates $y_1^{(k)},\ y_2^{(k)},\ \ldots,\ y_n^{(k)}$ of the vector $\mathbf{y}^{(k)}$ are successively computed from the formulas

$$\left.\begin{array}{l} y_i^{(1)}=\sum\limits_{j=1}^{n} a_{ij}y_i^{(0)}, \\ y_i^{(2)}=\sum\limits_{j=1}^{n} a_{ij}y_i^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_i^{(n)}=\sum\limits_{j=1}^{n} a_{ij}y_j^{(n-1)} \qquad (i=1,\ 2,\ \ldots,\ n) \end{array}\right\} \tag{7}$$

Thus, determining the coefficients p_j of the characteristic polynomial (1) by the Krylov method reduces to solving the linear system of equations (6), whose coefficients are computed from the formulas (7); note that the coordinates of the initial vector

$$\mathbf{y}^{(0)}=\begin{bmatrix} y_1^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix}$$

are arbitrary. If system (6) has a unique solution, then its roots $p_1,\ p_2,\ \ldots,\ p_n$ are the coefficients of the characteristic polynomial (1). This solution can be found, for example, by the Gaussian method (Sec. 8.3). If system (6) does not have a unique solution, the problem is complicated [1]. In this case it is advisable to change the initial vector.

Example. Use Krylov's method to find the characteristic polynomial of the matrix (see Sec. 12.3)

$$A=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}$$

Solution. We choose the initial vector

$$y^{(0)}=\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$$

Using formulas (7), we determine the coordinates of the vectors

$$y^{(k)}=A^k y^{(0)}\qquad (k=1,\ 2,\ 3,\ 4)$$

Thus; we have

$$y^{(1)}=Ay^{(0)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}=\begin{bmatrix}1\\2\\3\\4\end{bmatrix},$$

$$y^{(2)}=Ay^{(1)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}1\\2\\3\\4\end{bmatrix}=\begin{bmatrix}30\\22\\18\\20\end{bmatrix},$$

$$y^{(3)}=Ay^{(2)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}30\\22\\18\\20\end{bmatrix}=\begin{bmatrix}208\\178\\192\\242\end{bmatrix},$$

$$y^{(4)}=Ay^{(3)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}208\\178\\192\\242\end{bmatrix}=\begin{bmatrix}2108\\1704\\1656\\1992\end{bmatrix}$$

We set up the system (6):

$$\begin{bmatrix} y_1^{(3)} & y_1^{(2)} & y_1^{(1)} & y_1^{(0)} \\ y_2^{(3)} & y_2^{(2)} & y_2^{(1)} & y_2^{(0)} \\ y_3^{(3)} & y_3^{(2)} & y_3^{(1)} & y_3^{(0)} \\ y_4^{(3)} & y_4^{(2)} & y_4^{(1)} & y_4^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} y_1^{(4)} \\ y_2^{(4)} \\ y_3^{(4)} \\ y_4^{(4)} \end{bmatrix}$$

which in our case has the form

$$\begin{bmatrix} 208 & 30 & 1 & 1 \\ 178 & 22 & 2 & 0 \\ 192 & 18 & 3 & 0 \\ 242 & 20 & 4 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} 2108 \\ 1704 \\ 1656 \\ 1992 \end{bmatrix}$$

whence

$$\left.\begin{aligned} 208p_1 + 30p_2 + p_3 + p_4 &= -2108, \\ 178p_1 + 22p_2 + 2p_3 &= -1704, \\ 192p_1 + 18p_2 + 3p_3 &= -1656, \\ 242p_1 + 20p_2 + 4p_3 &= -1992 \end{aligned}\right\}$$

Solving this system, we obtain

$$p_1 = -4, \quad p_2 = -40, \quad p_3 = -56, \quad p_4 = -20$$

Hence

$$\det(\lambda E - A) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20$$

which coincides with the result obtained by the Danilevsky method (see Sec. 12.3).

127. COMPUTATION OF EIGENVECTORS BY THE KRYLOV METHOD

The method of A. N. Krylov makes it possible, in simple fashion, to find the corresponding eigenvectors [1].

For the sake of simplicity, we confine ourselves to the case when the characteristic polynomial

$$D(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

has distinct roots $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let us assume that the coefficients of the polynomial (1) and its roots have been determined. It is required to find the eigenvectors $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively.

Let $y^{(0)}$, $y^{(1)} = Ay^{(0)}$, $\ldots$, $y^{(n-1)} = A^{n-1}y^{(0)}$ be the vectors used in Krylov's method for finding the coefficients p_i $(i = 1, 2, \ldots, n)$. Decomposing the vector $y^{(0)}$ into the eigenvectors $x^{(i)}$ $(i = 1, 2, \ldots, n)$

we have

$$\boldsymbol{y}^{(0)}=c_1\boldsymbol{x}^{(1)}+c_2\boldsymbol{x}^{(2)}+\ldots+c_n\boldsymbol{x}^{(n)} \tag{2}$$

where c_i $(i=1, 2, \ldots, n)$ are certain numerical coefficients. From this, taking into consideration that

$$\begin{aligned} A\boldsymbol{x}^{(i)}&=\lambda_i\boldsymbol{x}^{(i)},\\ A^2\boldsymbol{x}^{(i)}&=\lambda_i^2\boldsymbol{x}^{(i)} \qquad (i=1, 2, \ldots, n)\\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned}$$

we get

$$\begin{aligned} \boldsymbol{y}^{(1)}&=c_1\lambda_1\boldsymbol{x}^{(1)}+c_2\lambda_2\boldsymbol{x}^{(2)}+\ldots+c_n\lambda_n\boldsymbol{x}^{(n)},\\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\\ \boldsymbol{y}^{(n-1)}&=c_1\lambda_1^{n-1}\boldsymbol{x}^{(1)}+c_2\lambda_2^{n-1}\boldsymbol{x}^{(2)}+\ldots+c_n\lambda_n^{n-1}\boldsymbol{x}^{(n)} \end{aligned} \tag{3}$$

Let

$$\varphi_i(\lambda)=\lambda^{n-1}+q_{1i}\lambda^{n-2}+\ldots+q_{n-1,i} \tag{4}$$

$(i=1, 2, \ldots, n)$ be an arbitrary set of polynomials. Forming a linear combination of the vectors $\boldsymbol{y}^{(n-1)}, \boldsymbol{y}^{(n-2)}, \ldots, \boldsymbol{y}^{(0)}$ with the coefficients $1, q_{1i}, \ldots, q_{n-1,i}$ we find, by relations (2) and (3),

$$\begin{aligned} &\boldsymbol{y}^{(n-1)}+q_{1i}\boldsymbol{y}^{(n-2)}+\ldots+q_{n-1,i}\boldsymbol{y}^{(0)}=\\ &\quad=c_1\varphi_i(\lambda_1)\,\boldsymbol{x}^{(1)}+c_2\varphi_i(\lambda_2)\,\boldsymbol{x}^{(2)}+\ldots+c_n\varphi_i(\lambda_n)\,\boldsymbol{x}^{(n)} \end{aligned} \tag{5}$$

If we put

$$\varphi_i(\lambda)=\frac{D(\lambda)}{\lambda-\lambda_i} \qquad (i=1, 2, \ldots, n) \tag{6}$$

then obviously

$$\varphi_i(\lambda_j)=0 \quad \text{for} \quad i\neq j$$

and

$$\varphi_i(\lambda_i)=D'(\lambda_i)\neq 0$$

Formula (5) then becomes

$$c_i\varphi_i(\lambda_i)\,\boldsymbol{x}^{(i)}=\boldsymbol{y}^{(n-1)}+q_{1i}\boldsymbol{y}^{(n-2)}+\ldots+q_{n-1,i}\boldsymbol{y}^{(0)} \qquad (i=1, 2, \ldots, n) \tag{7}$$

Thus, if $c_i\neq 0$, then the resulting linear combination of vectors $\boldsymbol{y}^{(n-1)}, \boldsymbol{y}^{(n-2)}, \ldots, \boldsymbol{y}^{(0)}$ yields the eigenvector $\boldsymbol{x}^{(i)}$ to within a numerical factor. The coefficients q_{ji} $(j=1, 2, \ldots, n-1)$ can easily be found from the *Horner scheme*

$$\left.\begin{aligned} q_{0i}&=1,\\ q_{ji}&=\lambda_i q_{j-1,\,i}+p_j \end{aligned}\right\}$$

12.8 LEVERRIER'S METHOD

This method of expanding a secular determinant is based on the Newtonian formulas [3] for the sums of powers of the roots of an algebraic equation.

Let

$$\det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

be the characteristic polynomial of a given matrix $A = [a_{ij}]$ and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the complete set of its roots, where each root is repeated as many times as its multiplicity.

Set

$$s_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k \qquad (k = 0, 1, 2, \ldots, n)$$

Then for $k \leqslant n$ the *Newtonian formulas* [3] hold true:

$$s_k + p_1 s_{k-1} + \ldots + p_{k-1}s_1 = -kp_k \qquad (k = 1, 2, \ldots, n) \tag{2}$$

whence

$$\left.\begin{aligned} p_1 &= -s_1, \\ p_2 &= -\frac{1}{2}(s_2 + p_1 s_1), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ p_n &= -\frac{1}{n}(s_n + p_1 s_{n-1} + \ldots + p_{n-1}s_1) \end{aligned}\right\} \tag{3}$$

If the sums $s_1, s_2, \ldots, s_n$ are known, then by means of formulas (3) we can, step by step, determine the coefficients $p_1, p_2, \ldots, p_n$ of the characteristic polynomial (1).

The sums $s_1, s_2, \ldots, s_n$ are computed as follows: for s_1 we have (Sec. 10.12)

$$s_1 = \lambda_1 + \lambda_2 + \ldots + \lambda_n = \operatorname{tr} A$$

that is,

$$s_1 = \sum_{i=1}^{n} a_{ii} \tag{4}$$

Then, as we know (Sec. 11.1), $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ are the eigenvalues of the matrix A^k. Therefore

$$s_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k = \operatorname{tr} A^k$$

That is, if

$$A^k = [a_{ij}^{(k)}]$$

then

$$s_k = \sum_{i=1}^{n} a_{ii}^{(k)} \tag{5}$$

The powers $A^k = A^{k-1}A$ are found by direct multiplication.

Thus, the scheme for expanding a secular determinant by the Leverrier method is extremely simple, namely: first compute A^k $(k=1, 2, \ldots, n)$, the powers of the given matrix A, then find the corresponding s_k, which are the sums of the elements of the principal diagonals of the matrices A^k and, finally, determine from formulas (3) the desired coefficients p_i $(i=1, 2, \ldots, n)$.

The Leverrier method is extremely laborious because one has to compute high powers of the given matrix. Its merit lies in the simple computational scheme and the absence of exceptional cases.

Example. Use the Leverrier method to expand the characteristic determinant of the matrix (see Sec. 12.3)

$$A=\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}$$

Solution. Form the powers A^k $(k=2, 3, 4)$ of the matrix A. We have

$$A^2=\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}=\begin{bmatrix}30 & 22 & 18 & 20\\ 22 & 18 & 16 & 18\\ 18 & 16 & 18 & 22\\ 20 & 18 & 22 & 30\end{bmatrix},$$

$$A^3=\begin{bmatrix}30 & 22 & 18 & 20\\ 20 & 18 & 16 & 18\\ 18 & 16 & 18 & 22\\ 20 & 18 & 22 & 30\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}=\begin{bmatrix}208 & 178 & 192 & 242\\ 178 & 148 & 154 & 192\\ 192 & 154 & 148 & 178\\ 242 & 192 & 178 & 208\end{bmatrix},$$

$$A^4=\begin{bmatrix}208 & 178 & 192 & 242\\ 178 & 148 & 154 & 192\\ 192 & 154 & 148 & 178\\ 242 & 192 & 178 & 208\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}=\begin{bmatrix}2108 & 1704 & 1656 & 1992\\ 1704 & 1388 & 1368 & 1656\\ 1656 & 1368 & 1388 & 1704\\ 1992 & 1656 & 1704 & 2108\end{bmatrix}$$

Note that it was not necessary to compute A^4 completely, it being sufficient to find only the elements of the principal diagonal of the matrix.

Whence

$$\begin{aligned}
s_1 &= \operatorname{tr} A = 1+1+1+1=4,\\
s_2 &= \operatorname{tr} A^2 = 30+18+18+30=96,\\
s_3 &= \operatorname{tr} A^3 = 208+148+148+208=712,\\
s_4 &= \operatorname{tr} A^4 = 2108+1388+1388+2108=6992
\end{aligned}$$

Hence, from formulas (3), we get

$$p_1 = -s_1 = -4,$$

$$p_2 = -\frac{1}{2}(s_2 + p_1 s_1) = -\frac{1}{2}(96 - 4\cdot 4) = -40,$$

$$p_3 = -\frac{1}{3}(s_3 + p_1 s_2 + p_2 s_1) = -\frac{1}{3}(712 - 4\cdot 96 - 40\cdot 4) = -56,$$

$$p_4 = -\frac{1}{4}(s_4 + p_1 s_3 + p_2 s_2 + p_3 s_1 = \\ = -\frac{1}{4}(6992 - 4\cdot 712 - 40\cdot 96 - 56\cdot 4) = -20$$

Thus, we obtain the already familiar result (see Sec. 12.3):

$$\begin{vmatrix} \lambda-1 & -2 & -3 & 4 \\ -2 & \lambda-1 & -2 & -3 \\ -3 & -2 & \lambda-1 & -2 \\ -4 & -3 & -2 & \lambda-1 \end{vmatrix} = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20$$

12.9 ON THE METHOD OF UNDETERMINED COEFFICIENTS

A secular determinant can also be expanded by finding a sufficiently large number of its numerical values.

Let

$$D(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

be the secular determinant of the matrix A, that is,

$$D(\lambda) = \det(\lambda E - A)$$

If in (1) we successively put $\lambda = 0, 1, 2, \ldots, n-1$, then for the coefficients p_i $(i = 1, 2, \ldots, n)$ we get the following system of linear equations

$$p_n = D(0),$$

$$\left.\begin{aligned} 1^n + p_1\cdot 1^{n-1} + \ldots + p_n &= D(1), \\ 2^n + p_1\cdot 2^{n-1} + \ldots + p_n &= D(2), \\ \ldots\ldots\ldots\ldots&\ldots\ldots \\ (n-1)^n + p_1(n-1)^{n-1} + \ldots + p_n &= D(n-1) \end{aligned}\right\} \tag{2}$$

whence

$$\left.\begin{aligned} &p_1 + p_2 + \ldots + p_{n-1} = D(1) - D(0) - 1, \\ &2^{n-1}p_1 + 2^{n-2}p_2 + \ldots + 2p_{n-1} = D(2) - D(0) - 2^n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &(n-1)^{n-1}p_1 + (n-1)^{n-2}p_2 + \ldots + (n-1)p_{n-1} = \\ &\qquad = D(n-1) - D(0) - (n-1) \end{aligned}\right\} \tag{3}$$

and

$$p_n = D(0) = \det(-A)$$

From system (3) we can determine the coefficients p_i ($i = 1, 2, \ldots, n$) of the characteristic polynomial (1).

Introducing the matrix

$$C_n = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 2^{n-1} & 2^{n-2} & \ldots & 2 \\ \cdot & \cdot & \cdot & \cdot \\ (n-1)^{n-1} & (n-1)^{n-2} & \ldots & n-1 \end{bmatrix}$$

and the vectors

$$D = \begin{bmatrix} D(1) - D(0) - 1^n \\ D(2) - D(0) - 2^n \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ D(n-1) - D(0) - (n-1)^n \end{bmatrix}, \quad P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \end{bmatrix}$$

we can write system (3) in the form of a matrix equation:

$$C_n P = D \tag{4}$$

whence

$$P = C_n^{-1} D \tag{5}$$

Note that the inverse matrix C_n^{-1} depends only on the order n of the secular determinant and can be found beforehand if one has to expand large numbers of secular determinants of the same order.

Thus, the use of this method reduces to computing the numerical determinants

$$D(k) = \det(kE - A) \qquad (k = 0, 1, 2, \ldots, n-1)$$

and to finding the solution of the standard linear system (4).

12.10 A COMPARISON OF DIFFERENT METHODS OF EXPANDING A SECULAR DETERMINANT

An indication of the relative effectiveness of various methods of expanding a secular determinant is given in Table 26 [4], which states the number of operations required by each method depending on the order of the determinant.

From this table it is seen that the best method for expanding secular determinants of order higher than fifth is, from the viewpoint of number of operations, the Danilevsky method.

12.11 FINDING THE NUMERICALLY LARGEST EIGENVALUE OF A MATRIX AND THE CORRESPONDING EIGENVECTOR

Suppose we have the characteristic equation

$$\det(A-\lambda E)=0$$

The roots of this equation $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of the matrix A. Suppose that to them correspond the linearly independent eigenvectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$. We will now give certain iteration methods for computing the numerically largest eigenvalue of matrix A that do not require expanding its secular determinant.

Case 1. Among the eigenvalues of matrix A there is one which is largest in modulus. For the sake of definiteness, let us assume that

$$|\lambda_1|>|\lambda_2|\geqslant|\lambda_3|\geqslant\ldots\geqslant|\lambda_n| \tag{1}$$

so that the numerically largest eigenvalue is the *first one.* It is obvious that for a real matrix the numerically largest eigenvalue λ_1 is real. We note that such is the case if the matrix A is real and its elements are positive (Sec. 10.16, Perron's theorem).

TABLE 26

THE NUMBER OF OPERATIONS USED BY VARIOUS METHODS IN EXPANDING A SECULAR DETERMINANT DEPENDING ON ITS ORDER

	Order									
Method	3		4		5		7		9	
	Multiplications-Divisions, M-D	Additions-Subtractions, A-S	M-D	A-S	M-D	A-S	M-D	A-S	M-D	A-S
Direct expansion	12	10	60	46	320	238	13 692	10 078	986 400	725 758
Danilevsky	14	12	42	36	92	80	282	252	632	576
Krylov	67	38	179	118	389	280	1287	1022	3209	2688
Leverrier	41	27	153	114	414	330	1791	1533	5228	4644
Undetermined coefficients	67	41	171	116	364	265	1189	945	2966	2481
Interpolation formula (see Sec. 14.23)	46	38	125	102	279	230	972	826	2525	2202

We now give an approximate method for computing the root λ_1. Take an arbitrary vector $\boldsymbol{y}$ and decompose it into the eigenvectors $\boldsymbol{x}^{(j)}$ of A:

$$\boldsymbol{y}=\sum_{j=1}^{n} c_j\boldsymbol{x}^{(j)}$$

where c_j $(j=1, 2, \ldots, n)$ are constant coefficients. Operating on vector $\boldsymbol{y}$ by the matrix A we have

$$A\boldsymbol{y}=\sum_{j=1}^{n} c_j A\boldsymbol{x}^{(j)}$$

whence, since $\boldsymbol{x}^{(j)}$ is an eigenvector of the transformation A, that is, $A\boldsymbol{x}^{(j)}=\lambda_j\boldsymbol{x}^{(j)}$, we get

$$A\boldsymbol{y}=\sum_{j=1}^{n} c_j\lambda_j\boldsymbol{x}^{(j)}$$

We call $A\boldsymbol{y}$ an *iteration* of the vector $\boldsymbol{y}$.

Forming the iterations $A\boldsymbol{y}$, $A^2\boldsymbol{y}$, $\ldots$, $A^m\boldsymbol{y}$ in succession, we find

$$A^m\boldsymbol{y}=\sum_{j=1}^{n} c_j\lambda_j^m\boldsymbol{x}^{(j)} \tag{2}$$

(*mth iteration*).

In the space $E_n=\{\boldsymbol{y}\}$ choose a basis $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\ldots$, $\boldsymbol{e}_n$ (not necessarily the unit basis). Let

$$A^m\boldsymbol{y}=\boldsymbol{y}^{(m)} \qquad (m=1, 2, 3, \ldots)$$

and

$$\boldsymbol{y}^{(m)}=\begin{bmatrix} y_1^{(m)} \\ \vdots \\ y_n^{(m)} \end{bmatrix}$$

where $y_i^{(m)}$ $(i=1, 2, \ldots, n)$ are the coordinates of the vector $\boldsymbol{y}^{(m)}$ in the chosen basis.

Decomposing the eigenvectors $\boldsymbol{x}^{(j)}$ into the vectors of the basis, we have

$$\boldsymbol{x}^{(j)}=\sum_{i=1}^{n} x_{ij}\boldsymbol{e}_i \tag{3}$$

whence, substituting (3) into (2), we obtain

$$\boldsymbol{y}^{(m)}=\sum_{j=1}^{n} c_j\lambda_j^m \sum_{i=1}^{n} x_{ij}\boldsymbol{e}_i$$

or, changing the order of summation,

$$\boldsymbol{y}^{(m)}=\sum_{i=1}^{n} \boldsymbol{e}_i \sum_{j=1}^{n} c_j x_{ij} \lambda_j^m \tag{4}$$

The coefficient of $\boldsymbol{e}_i$ is the ith coordinate of the vector $\boldsymbol{y}^{(m)}$. We can thus write

$$y_i^{(m)}=\sum_{j=1}^{n} c_j x_{ij} \lambda_j^m \tag{4$'$}$$

Similarly

$$y_i^{(m+1)}=\sum_{j=1}^{n} c_j x_{ij} \lambda_j^{m+1} \tag{4$''$}$$

Dividing the second sum by the first, we get

$$\frac{y_i^{(m+1)}}{y_i^{(m)}}=\frac{c_1 x_{i1} \lambda_1^{m+1}+\ldots+c_n x_{in} \lambda_n^{m+1}}{c_1 x_{i1} \lambda_1^m+\ldots+c_n x_{in} \lambda_n^m} \tag{5}$$

Suppose that $c_1 \neq 0$ and $x_{i1} \neq 0$. This can be achieved by appropriate choice of the initial vector $\boldsymbol{y}$ and the basis $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$. Transform expression (5) as follows:

$$\frac{y_i^{(m+1)}}{y_i^{(m)}}=\lambda_1 \frac{1+\frac{c_2 x_{i2}}{c_1 x_{i1}}\left(\frac{\lambda_2}{\lambda_1}\right)^{m+1}+\ldots+\frac{c_n x_{in}}{c_1 x_{i1}}\left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{1+\frac{c_2 x_{i2}}{c_1 x_{i1}}\left(\frac{\lambda_2}{\lambda_1}\right)^{m}+\ldots+\frac{c_n x_{in}}{c_1 x_{i1}}\left(\frac{\lambda_n}{\lambda_1}\right)^{m}}$$

From this, passing to the limit as $m \to \infty$ and taking into account inequality (1), we obtain

$$\lim_{m \to \infty} \frac{y_i^{(m+1)}}{y_i^{(m)}}=\lambda_1 \tag{6}$$

$\left(\text{since } \lim\limits_{m \to \infty}\left(\frac{\lambda_j}{\lambda_1}\right)^m=0 \text{ for } j>1\right)$ or, approximately,

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i=1, 2, \ldots, n) \tag{7}$$

and, more exactly,

$$\lambda_1=\frac{y_i^{(m+1)}}{y_i^{(m)}}+O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^m\right)$$

By taking a sufficiently large iteration number m, we can determine from (7) the largest (in modulus) root λ_1 of the characteristic equation of the given matrix A to any degree of accuracy. To find this root one can use any coordinate of the vector $\boldsymbol{y}^{(m)}$; in particular, one can take the arithmetic mean of the corresponding ratios.

Note 1. In exceptional cases when the initial vector $\boldsymbol{y}$ is not aptly chosen, formula (6) may not yield the required root or may even be meaningless, which is to say the limit of the ratio $\frac{y_i^{(m+1)}}{y_i^{(m)}}$ may not exist. This is readily seen from the "oscillating" values of the ratio. In such cases one should try a different initial vector.

Note 2. To accelerate the convergence of the iteration process (6) it is sometimes advantageous to form the sequence of matrices

$$\begin{aligned} &A^2 = A \cdot A, \\ &A^4 = A^2 \cdot A^2, \\ &A^8 = A^4 \cdot A^4, \\ &\cdots\cdots\cdots \\ &A^{2^k} = A^{2^{k-1}} \cdot A^{2^{k-1}} \end{aligned}$$

from which we find

$$\boldsymbol{y}^{(m)} = A^m \boldsymbol{y}$$

and

$$\boldsymbol{y}^{(m+1)} = A\boldsymbol{y}^{(m)}$$

where $m = 2^k$. Then we assume, as usual,

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i = 1, 2, \ldots, n)$$

The vector $\boldsymbol{y}^{(m)} = A^m \boldsymbol{y}$ is approximately the eigenvector of the matrix A corresponding to the eigenvalue λ_1.

Indeed, from formula (2) we have

$$A^m \boldsymbol{y} = c_1 \lambda_1^m \boldsymbol{x}^{(1)} + \sum_{j=2}^{n} c_j \lambda_j^m \boldsymbol{x}^{(j)}$$

where $\boldsymbol{x}^{(j)}$ $(j = 1, 2, \ldots, n)$ are the eigenvectors of matrix A.

From this,

$$A^m \boldsymbol{y} = c_1 \lambda_1^m \left\{ \boldsymbol{x}^{(1)} + \sum_{j=1}^{n} \frac{c_j}{c_1} \left(\frac{\lambda_j}{\lambda_1} \right)^m \boldsymbol{x}^{(j)} \right\}$$

Since $\left(\frac{\lambda_j}{\lambda_1} \right)^m \to 0$ as $m \to \infty$ $(j > 1)$, for a sufficiently large m we will have, to any degree of accuracy,

$$A^m \boldsymbol{y} \approx c_1 \lambda_1^m \boldsymbol{x}^{(1)}$$

Thus, $A^m\boldsymbol{y}$ differs from the eigenvector $\boldsymbol{x}^{(1)}$ solely by a numerical factor and, consequently, is also an eigenvector corresponding to the same eigenvalue λ_1.

Example. Find the largest eigenvalue of the matrix

$$A=\begin{bmatrix}4 & 1 & 0\\ 1 & 2 & 1\\ 0 & 1 & 1\end{bmatrix} \tag{8}$$

and the eigenvector corresponding to it.

Solution. Choose the initial vector

$$\boldsymbol{y}=\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$$

Form Table 27.

TABLE 27

COMPUTING THE FIRST EIGENVALUE

y	Ay	A^2y	A^3y	A^4y	A^5y	A^6y	A^7y	A^8y	A^9y	$A^{10}y$
1	5	24	111	504	2268	10161	45433	202833	905238	4038939
1	4	15	60	252	1089	4779	21141	93906	417987	1862460
1	2	6	21	81	333	1422	6201	27342	121248	539235

Stopping with the iterations $A^9\boldsymbol{y}=\boldsymbol{y}^{(9)}$ and $A^{10}\boldsymbol{y}=\boldsymbol{y}^{(10)}$, we get the values

$$\frac{y_1^{(10)}}{y_1^{(9)}}=\frac{4038939}{905238}=4.462,$$

$$\frac{y_2^{(10)}}{y_2^{(9)}}=\frac{1862460}{417987}=4.456,$$

$$\frac{y_3^{(10)}}{y_3^{(9)}}=\frac{539235}{121248}=4.447$$

Hence, we can take, approximately,

$$\lambda_1=\frac{1}{3}(4.462+4.456+4.447)=4.455\approx 4.46$$

For the first eigenvector of matrix A we can take

$$A^{10}\boldsymbol{y}=\begin{bmatrix}4038939\\ 1862460\\ 539235\end{bmatrix}$$

Normalizing it, we finally obtain

$$x^{(1)} = \begin{bmatrix} 0.90 \\ 0.42 \\ 0.12 \end{bmatrix}$$

Case 2. The largest, in modulus, eigenvalue of matrix A is multiple.

Let

$$\lambda_1 = \lambda_2 = \ldots = \lambda_s$$

and

$$|\lambda_1| > |\lambda_k| \quad \text{for} \quad k > s$$

From formula (5) we have

$$\frac{y_i^{(m+1)}}{y_i^{(m)}} = \frac{c_1 x_{i1} \lambda_1^{m+1} + \ldots + c_s x_{is} \lambda_1^{m+1} + c_{s+1} x_{i,\,s+1} \lambda_{s+1}^{m+1} + \ldots + c_n x_{in} \lambda_n^{m+1}}{c_1 x_{i1} \lambda_1^{m} + \ldots + c_s x_{is} \lambda_1^{m} + c_{s+1} x_{i,\,s+1} \lambda_{s+1}^{m} + \ldots + c_n x_{in} \lambda_n^{m}} =$$

$$= \lambda_1 \frac{c_1 x_{i1} + \ldots + c_s x_{is} + c_{s+1} x_{i,\,s+1} \left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m+1} + \ldots + c_n x_{in} \left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{c_1 x_{i1} + \ldots + c_s x_{is} + c_{s+1} x_{i,\,s+1} \left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m} + \ldots + c_n x_{in} \left(\frac{\lambda_n}{\lambda_1}\right)^{m}}$$

whence, if $c_1 x_{i1} + \ldots + c_s x_{is} \neq 0$ and taking into account that

$$\left(\frac{\lambda_k}{\lambda_1}\right)^m \to 0 \text{ as } m \to \infty \quad \text{and} \quad k > s$$

we get

$$\lim_{m \to \infty} \frac{y_i^{(m+1)}}{y_i^{(m)}} = \lambda_1 \qquad (i = 1,\ 2,\ \ldots,\ n)$$

or, more exactly,

$$\lambda_1 = \frac{y_i^{(m+1)}}{y_i^{(m)}} + O\left(\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^m\right).$$

Thus, the foregoing method for computing λ_1 is applicable in this case as well.

As before,

$$y^{(m)} = A^m y$$

is one of the approximate eigenvectors of matrix A corresponding to the eigenvalue λ_1. Generally speaking, by changing the initial vector y, we obtain a different linearly independent vector of matrix A. Note that in this case there is no guarantee that our technique will determine the entire set of linearly independent eigenvectors of A for the eigenvalue λ_1.

For Cases 1 and 2 we can offer a faster iteration process for finding the numerically largest eigenvalue λ_1 of matrix A, namely:

form the sequence of matrices

$$A,\ A^2,\ A^4,\ A^8,\ \dots,\ A^{2^k}$$

As we know (see Sec. 10.12),

$$\sum_{i=1}^{n} \lambda_i = \operatorname{tr} A$$

Similarly,

$$\sum_{i=1}^{n} \lambda_i^m = \operatorname{tr} A^m$$

where $m = 2^k$. Confining ourselves to Case 1 for the sake of simplicity, we have

$$\lambda_1^m + \lambda_2^m + \dots + \lambda_n^m = \lambda_1^m \left[1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m + \dots + \left(\frac{\lambda_n}{\lambda_1}\right)^m\right] = \operatorname{tr} A^m$$

whence

$$|\lambda_1| \left[1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m + \dots + \left(\frac{\lambda_n}{\lambda_1}\right)^m\right]^{\frac{1}{m}} = \sqrt[m]{\operatorname{tr} A^m}$$

As $m \longrightarrow \infty$ we get

$$|\lambda_1| = \lim_{m \to \infty} \sqrt[m]{\operatorname{tr} A^m}$$

or

$$|\lambda_1| \approx \sqrt[m]{\operatorname{tr} A^m}$$

where m is sufficiently great ($m \gg n$).

In order to avoid extraction of high-degree roots, we can find

$$A^{m+1} = A^m A$$

Then

$$\lambda_1^{m+1} + \lambda_2^{m+1} + \dots + \lambda_n^{m+1} = \operatorname{tr} A^{m+1}$$

and

$$\lambda_1^m + \lambda_2^m + \dots + \lambda_n^m = \operatorname{tr} A^m$$

whence, taking into account the smallness of $|\lambda_2|, \dots, |\lambda_n|$ as compared with $|\lambda_1|$, we obtain

$$\lambda_1 \approx \operatorname{tr} A^{m+1} / \operatorname{tr} A^m$$

12.12 THE METHOD OF SCALAR PRODUCTS FOR FINDING THE FIRST EIGENVALUE OF A REAL MATRIX

A somewhat different iteration process (and at times a more advantageous one) can be given for finding the first eigenvalue λ_1 of a real matrix A. This method is based on the formation of scalar products

$$(A^k \boldsymbol{y}_0,\ A'^k \boldsymbol{y}_0) \quad \text{and} \quad (A^{k-1} \boldsymbol{y}_0,\ A'^k \boldsymbol{y}_0)$$

($k=1, 2, \ldots$), where A' is the transpose of A and $\boldsymbol{y}_0$ is the initial vector chosen in some manner.

Let us now take up the method itself.

Suppose A is a real matrix and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are its eigenvalues which are assumed to be distinct, and

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|.$$

We take some nonzero vector $\boldsymbol{y}_0$ and with the aid of matrix A construct a sequence of iterations

$$\boldsymbol{y}_k = A\boldsymbol{y}_{k-1} \qquad (k=1, 2, \ldots) \tag{1}$$

For the vector $\boldsymbol{y}_0$ we also form, using the transpose A', a second sequence of iterations

$$\boldsymbol{y}'_k = A'\boldsymbol{y}'_{k-1} \qquad (k=1, 2, \ldots) \tag{2}$$

where $\boldsymbol{y}'_0 = \boldsymbol{y}_0$.

By Theorem 1 of Sec. 10.16, in the space E_n we choose two proper bases $\{\boldsymbol{x}_j\}$ and $\{\boldsymbol{x}'_j\}$ for the matrices A and A', respectively, satisfying the conditions of biorthonormalization:

$$(\boldsymbol{x}_i,\ \boldsymbol{x}'_j) = \delta_{ij} \tag{3}$$

where $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$ and $A'\boldsymbol{x}'_j = \lambda^*_j \boldsymbol{x}'_j$ ($i, j = 1, 2, \ldots, n$). Denote the coordinates of the vector $\boldsymbol{y}_0$ in the basis $\{\boldsymbol{x}_j\}$ by $a_1, \ldots, a_n$, and in the basis $\{\boldsymbol{x}'_j\}$ by $b_1, \ldots, b_n$, that is,

$$\boldsymbol{y}_0 = a_1\boldsymbol{x}_1 + \ldots + a_n\boldsymbol{x}_n \quad \text{and} \quad \boldsymbol{y}_0 = b_1\boldsymbol{x}'_1 + \ldots + b_n\boldsymbol{x}'_n$$

whence

$$\boldsymbol{y}_k = A^k\boldsymbol{y}_0 = \sum_{j=1}^{n} a_j \lambda_i^k \boldsymbol{x}_j \tag{4}$$

and

$$\boldsymbol{y}'_k = A'^k\boldsymbol{y}_0 = \sum_{j=1}^{n} b_j \lambda_j^{*k} \boldsymbol{x}'_j \qquad (k=1, 2, \ldots) \tag{4'}$$

Form the scalar product

$$(\boldsymbol{y}_k,\ \boldsymbol{y}'_k) = (A^k\boldsymbol{y}_0,\ A'^k\boldsymbol{y}_0) = (\boldsymbol{y}_0,\ A'^{2k}\boldsymbol{y}_0) = \left(\sum_{i=1}^{n} a_i\boldsymbol{x}_i,\ \sum_{j=1}^{n} b_j\lambda_j^{*2k}\boldsymbol{x}'_j\right)$$

From this, by virtue of the orthonormalization condition, we find

$$(\boldsymbol{y}_k,\ \boldsymbol{y}'_k) = \sum_{j=1}^{n} a_j b^*_j \lambda_j^{2k} = a_1 b^*_1 \lambda_1^{2k} + a_2 b^*_2 \lambda_2^{2k} + \ldots + a_n b^*_n \lambda_n^{2k} \tag{5}$$

Similarly

$$(y_{k-1}, y'_k) = a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1} \tag{6}$$

Hence, for $a_1 b_1^* \neq 0$ we have

$$\frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{a_1 b_1^* \lambda_1^{2k} + a_2 b_2^* \lambda_2^{2k} + \ldots + a_n b_n^* \lambda_n^{2k}}{a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1}} = \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)$$

Thus

$$\lambda_1 \approx \frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{(A^k y_0, A'^k y_0)}{(A^{k-1} y_0, A'^k y_0)} \tag{7}$$

This method is especially convenient for a symmetric matrix A, since then $A' = A$, and we simply have

$$\lambda_1 \approx \frac{(A^k y_0, A^k y_0)}{(A^{k-1} y_0, A^k y_0)} \tag{8}$$

and so we only have to construct one sequence $y_k = A^k y_0$ $(k = 1, 2, \ldots)$.

Example. Use the method of scalar products to find the largest eigenvalue of the matrix (see Sec. 12.11)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Solution. Since the matrix A is symmetric, it suffices to construct only one sequence of iterations $A^k y_0$ $(k = 1, 2, \ldots)$. Taking

$$y_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

for the initial vector we can use the results of Table 27. For instance, when $k = 5$ and $k = 6$ we have

$$A^5 y_0 = \begin{bmatrix} 2268 \\ 1089 \\ 333 \end{bmatrix} \quad \text{and} \quad A^6 y_0 = \begin{bmatrix} 10161 \\ 4779 \\ 1422 \end{bmatrix}$$

whence

$$(A^5 y_0, A^6 y_0) = 2268 \cdot 10{,}161 + 1089 \cdot 4779 + 333 \cdot 1422 = 28{,}723{,}005$$

and

$$(A^6 y_0, A^6 y_0) = 10{,}161^2 + 4779^2 + 1422^2 = 128{,}106{,}846$$

And so

$$\lambda_1 \approx \frac{(A^6 y_0,\ A^6 y_0)}{(A^5 y_0,\ A^6 y_0)} = \frac{128{,}106{,}846}{28{,}723{,}005} = 4.46$$

which coincides (within the digits written) with the value found earlier with the aid of $A^{10} y_0$ (see Sec. 12.11).

Note. The method for finding the numerically largest root of a characteristic equation (Sec. 12.11) may be used to find the numerically largest root of an algebraic equation:

$$x^n + p_1 x^{n-1} + \ldots + p_n = 0 \tag{9}$$

Indeed, equation (9), as may be readily verified directly, is the secular equation of the matrix (cf. Sec. 12.3, Frobenius matrix)

$$P = \begin{bmatrix} -p_1 & -p_2 & \ldots & -p_{n-1} & -p_n \\ 1 & 0 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}$$

which means that equation (9) is equivalent to the equation

$$\det (xP - E) = 0$$

If (9) does not have zero roots, then, in analogous fashion, we can determine the smallest (in modulus) root of the equation, namely for $p_n \neq 0$, assuming $\frac{1}{x} = y$, we obtain

$$y^n + \frac{p_{n-1}}{p_n} y^n + \ldots + \frac{1}{p_n} = 0 \tag{10}$$

The reciprocal of the numerically largest root of (10) will obviously give us the numerically smallest root of equation (9).

12.13 FINDING THE SECOND EIGENVALUE OF A MATRIX AND THE SECOND EIGENVECTOR

Suppose the eigenvalues λ_j $(j = 1, 2, \ldots, n)$ of matrix A are such that

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n| \tag{1}$$

that is, there are two distinct numerically largest eigenvalues λ_1 and λ_2 of matrix A. In this case, we can use a device similar to the one discussed above (Sec. 12.11) to approximate the second eigenvalue λ_2 and the eigenvector $x^{(2)}$ corresponding to it.

From formula (2) of Sec. 12.11 we have

$$A^m y = c_1 \lambda_1^m x^{(1)} + c_2 \lambda_2^m x^{(2)} + \ldots + c_n \lambda_n^m x^{(n)} \tag{2}$$

and

$$A^{m+1}\boldsymbol{y} = c_1\lambda_1^{m+1}\boldsymbol{x}^{(1)} + c_2\lambda_2^{m+1}\boldsymbol{x}^{(2)} + \dots + c_n\lambda_n^{m+1}\boldsymbol{x}^{(n)} \tag{3}$$

Eliminate from (2) and (3) the terms containing λ_1. To do this, subtract from (3) the equation (2) multiplied by λ_1 to get

$$A^{m+1}\boldsymbol{y} - \lambda_1 A^m\boldsymbol{y} = c_2\lambda_2^m(\lambda_2 - \lambda_1)\boldsymbol{x}^{(2)} + \dots + c_n\lambda_n^m(\lambda_n - \lambda_1)\boldsymbol{x}^{(n)} \tag{4}$$

For the sake of brevity, we introduce the notation

$$\Delta_\lambda A^m\boldsymbol{y} = A^{m+1}\boldsymbol{y} - \lambda A^m\boldsymbol{y} \tag{5}$$

We will call expression (5) the *λ-difference* of $A^m\boldsymbol{y}$. If $c_2 \neq 0$, then, clearly, the first term in the right member of (4) is its principal term as $m \to \infty$ and we have the approximate equation

$$\Delta_{\lambda_1} A^m\boldsymbol{y} \approx c_2\lambda_2^m(\lambda_2 - \lambda_1)\boldsymbol{x}^{(2)} \tag{6}$$

whence

$$\Delta_{\lambda_1} A^{m-1}\boldsymbol{y} \approx c_2\lambda_2^{m-1}(\lambda_2 - \lambda_1)\boldsymbol{x}^{(2)} \tag{7}$$

Let

$$A^m\boldsymbol{y} = \boldsymbol{y}^{(m)} = \begin{bmatrix} y_1^{(m)} \\ y_2^{(m)} \\ \vdots \\ y_n^{(m)} \end{bmatrix}$$

From formulas (6) and (7) we derive

$$\lambda_2 \approx \frac{\Delta_{\lambda_1} y_i^{(m)}}{\Delta_{\lambda_1} y_i^{(m-1)}} = \frac{y_i^{(m+1)} - \lambda_1 y_i^{(m)}}{y_i^{(m)} - \lambda_1 y_i^{(m-1)}} \quad (i = 1, 2, \dots, n) \tag{8}$$

Using formula (8) we can approximately compute the second eigenvalue λ_2. Note that in practical situations it is sometimes better (because of loss of accuracy when subtracting nearly equal numbers) to take a smaller iteration number k for determining λ_2 than the iteration number m for determining λ_1; in other words, it is advisable to set

$$\lambda_2 \approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} \qquad (k < m) \tag{9}$$

where k is the smallest of the numbers for which λ_2 begins to dominate the subsequent eigenvalues. Generally speaking, formula (9) yields rough values of λ_2. It will be seen that if the moduli of all the eigenvalues are distinct, then by means of formulas similar to (9) one can also compute the remaining eigenvalues of a given matrix. However, the results of the computations will be still less reliable.

As for the eigenvector $x^{(2)}$, we can, as follows from (6), put

$$x^{(2)} \approx \Delta_{\lambda_1} y^{(k)} \tag{10}$$

There is also an extension of this method to the case of multiple roots of a characteristic equation [1].

Example. Determine the subsequent eigenvalues and eigenvectors of the matrix (see Sec. 12.11, Example)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Solution. To find the second eigenvalue we take $k=8$. We have (see Table 27)

A^7y	A^8y	A^9y
45433 21141 6201	202833 93906 27342	905238 417987 121248

Form the λ-differences using the formula

$$\Delta_{\lambda_1} y_i^j = y_i^{(j+1)} - \lambda_1 y_i^{(j)} \qquad (i=1,\ 2,\ 3)$$

where $y^{(j)} = A^j y$. For each of the columns, λ_1 assumes a value: $\lambda_1 = 4.462$, $\lambda_1 = 4.456$, $\lambda_1 = 4.447$ (Table 28).

TABLE 28

COMPUTING THE SECOND EIGENVALUE

A^8y	$\lambda_1 A^7y$	$\Delta_{\lambda_1} A^7y$	A^9y	$\lambda_1 A^8y$	$\Delta_{\lambda_1} A^8y$
202833	202722	111	905238	905041	197
93906	94204	—298	417987	418445	—458
27342	27576	—234	121248	121590	—342

From this we get

$$\frac{\Delta_{\lambda_1} y_1^{(8)}}{\Delta_{\lambda_1} y_1^{(7)}} = \frac{197}{111} = 1.78, \qquad \frac{\Delta_{\lambda_1} y_2^{(8)}}{\Delta_{\lambda_1} y_2^{(7)}} = \frac{-458}{-298} = 1.54,$$

$$\frac{\Delta_{\lambda_1} y_3^{(8)}}{\Delta_{\lambda_1} y_3^{(7)}} = \frac{-342}{-234} = 1.46.$$

And so, we have approximately,

$$\lambda_2 = \frac{1}{3}(1.78 + 1.54 + 1.46) \approx 1.59$$

For the second eigenvector we can take

$$\Delta_{\lambda_1} A^8 \boldsymbol{y} = \begin{bmatrix} 197 \\ -458 \\ -342 \end{bmatrix}$$

Normalizing this vector, we obtain

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 0.33 \\ -0.76 \\ -0.56 \end{bmatrix}$$

Since matrix A is symmetric, the vectors $\boldsymbol{x}^{(1)}$ (Sec. 12.11) and $\boldsymbol{x}^{(2)}$ must be mutually orthogonal. Verification yields

$$(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = 0.90 \cdot 0.33 + 0.42 \cdot (-0.76) + 0.12 \cdot (-0.56) = 0.09$$

whence the angle $(x^{(1)}, x^{(2)}) = 85°$, which is rather inaccurate.

The third eigenvalue λ_3 is found from the trace of A:

$$\lambda_1 + \lambda_2 + \lambda_3 = \operatorname{tr} A = 4 + 2 + 1 = 7$$

whence

$$\lambda_3 = 7 - 4.46 - 1.56 \approx 0.95$$

The eigenvector

$$\boldsymbol{x}^{(3)} = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix}$$

may be computed from the orthogonality conditions:

$$\left.\begin{aligned} 0.90x_1^{(3)} + 0.42x_2^{(3)} + 0.12x_3^{(3)} &= 0, \\ 0.33x_1^{(3)} + (-0.76)\,x_2^{(3)} + (-0.56)\,x_3^{(3)} &= 0 \end{aligned}\right\}$$

whence

$$\frac{x_1^{(3)}}{\begin{vmatrix} 0.42 & 0.12 \\ -0.76 & -0.56 \end{vmatrix}} = \frac{x_2^{(3)}}{\begin{vmatrix} 0.12 & 0.90 \\ -0.56 & 0.33 \end{vmatrix}} = \frac{x_3^{(3)}}{\begin{vmatrix} 0.90 & 0.42 \\ 0.33 & -0.76 \end{vmatrix}}$$

or

$$\frac{x_1^{(3)}}{-0.144} = \frac{x_2^{(3)}}{0.539} = \frac{x_3^{(3)}}{-0.818}$$

Normalizing, we finally get

$$x^{(3)}=\begin{bmatrix}-0.14\\0.53\\-0.81\end{bmatrix}$$

12.14 THE METHOD OF EXHAUSTION

There is yet another method, called the *method of exhaustion* [1], for determining the second eigenvalue of a matrix and the eigenvector belonging to it.

Suppose matrix $A=[a_{ij}]$ is real, has distinct eigenvalues λ_1, $\lambda_2, \ldots, \lambda_n$ and

$$|\lambda_1|>|\lambda_2|>|\lambda_3|\geqslant\ldots\geqslant|\lambda_n|$$

Besides A, we consider the matrix

$$A_1=A-\lambda_1 X_1X_1' \tag{1}$$

where λ_1 is the first eigenvalue of A,

$$X_1=\begin{bmatrix}x_{11}\\x_{21}\\\vdots\\x_{n1}\end{bmatrix}$$

is the corresponding eigenvector of A regarded as a column matrix, and

$$X_1'=[x_{11}'\ x_{21}'\ \ldots\ x_{n1}']$$

is the eigenvector, corresponding to λ_1, of the transpose A' which is regarded as a row matrix; the vectors X_1 and X_1' are normalized so that their scalar product is equal to unity:

$$(X_1,\ X_1'^*)=X_1'X_1=\sum_{j=1}^{n}x_{j1}x_{j1}'=1 \tag{2}$$

We assume λ_1 and X_1 and X_1' to be known.

In expanded form, the matrix A_1 is written as

$$A_1=\begin{bmatrix}a_{11}&a_{12}&\ldots&a_{1n}\\a_{21}&a_{22}&\ldots&a_{2n}\\\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&\ldots&a_{nn}\end{bmatrix}-\lambda_1\begin{bmatrix}x_{11}\\x_{21}\\\vdots\\x_{n1}\end{bmatrix}[x_{11}'x_{21}'\ \ldots\ x_{n1}']=$$

$$= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} - \lambda_1 \begin{bmatrix} x_{11}x'_{11} & x_{11}x'_{21} & \dots & x_{11}x'_{n1} \\ x_{21}x'_{11} & x_{21}x'_{21} & \dots & x_{21}x'_{n1} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1}x'_{11} & x_{n1}x'_{21} & \dots & x_{n1}x'_{n1} \end{bmatrix} \qquad (1')$$

We will prove that the eigenvectors X_j $(j = 1, 2, \dots, n)$ of matrix A are also the eigenvectors of A_1, and the corresponding eigenvalues are preserved, with the exception of λ_1, in place of which a zero eigenvalue appears.

Indeed, using the associativity of matrix multiplication and the normalization condition (2), we have

$$A_1X_1 = AX_1 - \lambda_1 (X_1X'_1) X_1 = \lambda_1 X_1 - \lambda_1 X_1 (X'_1X_1) = \lambda_1 X_1 - \lambda_1 X_1 = 0$$

or

$$A_1X_1 = 0X_1$$

and, hence, zero is an eigenvalue of the matrix A_1.

Then, for $j > 1$, taking into account that

$$(X_j, X_1'^{*}) = X'_1X_j = 0 \qquad (j = 2, \dots, n)$$

(see Sec. 10.16, Theorem 1), we obtain

$$A_1X_j = AX_j - \lambda_1 (X_1X'_1) X_j = \lambda_1 X_j - \lambda_1 X_1 (X'_1X_j) = \lambda_j X_j$$
$$(j = 2, \dots, n)$$

Thus, λ_2 is the numerically largest eigenvalue of A_1, and so we can make use of the earlier indicated methods (Secs. 12.11 and 12.12) to determine λ_2 and the associated eigenvector X_2. This technique is called the *method of exhaustion*. For example, proceeding from an arbitrary vector $\boldsymbol{y}_0$, we can compute λ_2 from the formula

$$\lambda_2 \approx \frac{(A_1^m \boldsymbol{y}_0)_i}{(A_1^{m-1} \boldsymbol{y}_0)_i} \qquad (i = 1, 2, \dots, n)$$

Also

$$X_2 \approx c A_1^m \boldsymbol{y}_0 \quad (c \neq 0)$$

We will now show that to find the iterations $A_1^m \boldsymbol{y}_0$ $(m = 1, 2, \dots)$ one can use the formula

$$A_1^m \boldsymbol{y}_0 = A^m \boldsymbol{y}_0 - \lambda_1^m X_1 X'_1 \boldsymbol{y}_0 \qquad (3)$$

which dispenses with direct iteration of the matrix A_1.

Indeed, let the eigenvectors X_j and X'_j $(j = 1, 2, \dots, n)$ of matrix A and its transpose A' satisfy the conditions of biorthonormalization (Sec. 10.16, Theorem 2)

$$X'_kX_j = \delta_{jk}$$

where δ_{jk} is the Kronecker delta. We then have the bilinear expansion of A

$$A=\lambda_1X_1X_1'+\lambda_2X_2X_2'+\ldots+\lambda_nX_nX_n' \qquad (4)$$

whence

$$A_1=A-\lambda_1X_1X_1'=\lambda_2X_2X_2'+\ldots+\lambda_nX_nX_n' \qquad (5)$$

Since

$$A^mX_j=\lambda_j^mX_j \qquad (j=1,\ 2,\ \ldots,\ n)$$

then, by premultiplying (4) by A^{m-1}, we get

$$\begin{aligned}A^m&=A^mX_1X_1'+A^mX_2X_2'+\ldots+A^mX_nX_n'=\\&=\lambda_1^mX_1X_1'+\lambda_2^mX_2X_2'+\ldots+\lambda_n^mX_nX_n'\end{aligned} \qquad (6)$$

Similarly, taking into consideration that

$$A_1^mX_1=A_1^{m-1}(A_1X_1)=0$$

and

$$A_1^mX_j=\lambda_j^mX_j \qquad (j=2,\ 3,\ \ldots,\ n)$$

we obtain

$$A_1^m=A_1^mX_2X_2'+\ldots+A_1^mX_nX_n'=\lambda_2^mX_2X_2'+\ldots+\lambda_n^mX_nX_n' \qquad (7)$$

after premultiplying (5) by A_1^{m-1}. From formulas (6) and (7) follows

$$A_1^m=A^m-\lambda_1^mX_1X_1'$$

which is equivalent to relation (3).

12.15 FINDING THE EIGENVALUES AND EIGENVECTORS OF A POSITIVE DEFINITE SYMMETRIC MATRIX

Here we give an iteration method for finding simultaneously the eigenvalues and eigenvectors of a positive definite matrix [5].

As we know, (Sec. 10.15), if a real matrix

$$A=[a_{ij}]$$

is symmetric and positive definite, then

(1) the roots $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_n$ of its characteristic equation

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{vmatrix}=0 \qquad (1)$$

are real and positive:

(2) the eigenvectors

$$\boldsymbol{x}^{(j)} = \begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_n^{(j)} \end{bmatrix} \qquad (j = 1, 2, \ldots, n)$$

can be taken real and satisfy the orthogonality conditions

$$\sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} = 0 \quad \text{for} \quad j \neq k \tag{2}$$

Let us write the system for determining the eigenvector $\boldsymbol{x}^{(1)}$:

$$\left.\begin{array}{l} (a_{11}-\lambda_1)\, x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)} = 0, \\ a_{21}x_1^{(1)} + (a_{22}-\lambda_1)\, x_2^{(1)} + \ldots + a_{2n}x_n^{(1)} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + (a_{nn}-\lambda_1)\, x_n^{(1)} = 0 \end{array}\right\}$$

or

$$\left.\begin{array}{l} x_1^{(1)} = \dfrac{1}{\lambda_1}\left(a_{11}x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)}\right), \\ x_2^{(1)} = \dfrac{1}{\lambda_1}\left(a_{21}x_1^{(1)} + a_{22}x_2^{(1)} + \ldots + a_{2n}x_n^{(1)}\right), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n-1}^{(1)} = \dfrac{1}{\lambda_1}\left(a_{n-1,\,1}x_1^{(1)} + a_{n-1,\,2}x_2^{(1)} + \ldots + a_{n-1,\,n}x_n^{(1)}\right) \\ \lambda_1 = \dfrac{1}{x_n^{(1)}}\left(a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + a_{nn}x_{(n)}^{(1)}\right) \end{array}\right\} \tag{3}$$

Since the coordinates of the eigenvectors are determined to within the proportionality factor, one of them is arbitrary; for example, except for a special case, we can put $x_n^{(1)} = 1$. Generally speaking, system (3) can be solved by the method of iteration [5] by choosing suitable initial values $x_i^{(1,0)}$, $\lambda_1^{(0)}$ and assuming

$$x_i^{(1,\,k+1)} = \frac{1}{\lambda_1^{(k)}}\left(\sum_{j=1}^{n-1} a_{ij}x_j^{(1,\,k)} + a_{in}\right) \qquad (i = 1, 2, \ldots, n-1),$$

$$\lambda_1^{(k+1)} = \sum_{j=1}^{n-1} a_{nj}x_j^{(1,\,k+1)} + a_{nn} \qquad (k = 0, 1, 2, \ldots)$$

It is also possible to take advantage of the Seidel process. Thus, we find the first root of the characteristic equation (1)

$$\lambda_1 \approx \lambda_1^{(k)} \tag{4}$$

and the first eigenvector

$$\boldsymbol{x}^{(1)} \approx \begin{bmatrix} x_1^{(1,k)} \\ \vdots \\ x_{n-1}^{(1,k)} \\ 1 \end{bmatrix}$$

To determine the second root λ_2 of equation (1) and the second eigenvector $\boldsymbol{x}^{(2)}$, write the appropriate system of equations:

$$\lambda_2 x_i^{(2)} = \sum_{j=1}^{n} a_{ij} x_j^{(2)} \qquad (i = 1, 2, \ldots, n) \tag{5}$$

We eliminate from the orthogonality relation

$$\sum_{j=1}^{n} x_j^{(1)} x_j^{(2)} = 0 \tag{6}$$

one of the unknowns $x_j^{(2)}$, say, $x_n^{(2)}$. Then system (5) is replaced by the equivalent system

$$\left.\begin{aligned} x_i^{(2)} &= \frac{1}{\lambda_2} \sum_{j=1}^{n-1} a_{ij}^{(2)} x_j^{(2)} \qquad (i = 1, 2, \ldots, n-2), \\ \lambda_2 &= \frac{1}{x_{n-1}^{(2)}} \sum_{j=1}^{n-1} a_{n-1,\,j}^{(2)} x_j^{(2)} \end{aligned}\right\} \tag{7}$$

Setting $x_{n-1}^{(2)} = 1$, we solve (7) by the method of iteration, and thus find the second root λ_2 of the characteristic equation (1) and the eigenvector $\boldsymbol{x}^{(2)}$; the nth coordinate of this vector is determined from the orthogonality condition (6). The remaining roots λ_j $(j = 3, \ldots, n)$ of (1) and the corresponding eigenvectors $\boldsymbol{x}^{(j)}$ are found in similar fashion.

We do not consider any exceptional cases that may arise in the use of this method.

Example. For the following matrix find the roots λ_j of the characteristic equation and the eigenvectors $\boldsymbol{x}^{(j)}$ [5]:

$$A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{bmatrix}.$$

Solution. The matrix A is symmetric and positive definite since

$$\Delta_1 = 4 > 0,$$
$$\Delta_2 = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix} = 16 > 0,$$
$$\Delta_3 = \det\ A = 80 > 0$$

The associated system is of the form

$$\left.\begin{aligned} \lambda_j\ x_1^{(j)} &= 4x_1^{(j)} + 2x_2^{(j)} + 2x_3^{(j)} \\ \lambda_j\ x_2^{(j)} &= 2x_1^{(j)} + 5x_2^{(j)} + x_3^{(j)} \\ \lambda_j\ x_3^{(j)} &= 2x_1^{(j)} + x_2^{(j)} + 6x_3^{(j)} \end{aligned}\right\} \quad (j = 1,\ 2,\ 3) \qquad (8)$$

Setting $j = 1$ and $x_3^{(1)} = 1$, we get

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{\lambda_1}\ (4x_1^{(1)} + 2x_2^{(1)} + 2), \\ x_2^{(1)} &= \frac{1}{\lambda_1}\ (2x_1^{(1)} + 5x_2^{(1)} + 1), \\ \lambda_1 &= 2x_1^{(1)} + x_2^{(1)} + 6 \end{aligned}\right\} \qquad (9)$$

We solve (9) by the method of iteration, choosing the initial values

$$x_1^{(1,\ 0)} = 1 \quad \text{and} \quad x_2^{(1,\ 0)} = 1$$

Then we get $\lambda_1^{(0)} = 9$ from the last equation of system (9). The computations are arranged in Table 29.

TABLE 29

USING THE ITERATION METHOD TO COMPUTE THE EIGENVALUES AND EIGENVECTORS OF A MATRIX WHICH CORRESPOND TO THE FIRST ROOT OF THE CHARACTERISTIC EQUATION

k	$x_1^{(1k)}$	$x_2^{(1k)}$	$x_3^{(1k)}$	$\lambda_1^{(k)}$
0	1	1	1	9
1	0.89	0.89	1	8.67
2	0.85	0.83	1	8.53
3	0.83	0.80	1	8.46
4	0.81	0.78	1	8.40
5	0.805	0.770	1	8.38
6	0.806	0.771	1	8.383
7	0.807	0.771	1	8.385
8	0.8074	0.7715	1	8.3863
9	0.8076	0.7717	1	8.3869
10	0.8076	0.7719	1	8.3871
11	0.8077	0.7720	1	8.3874

We can take

$$\lambda_1 = 8.3874$$

and

$$x^{(1)} = \begin{bmatrix} 0.8077 \\ 0.7720 \\ 1 \end{bmatrix}$$

Now put $j=2$ in system (8). From the orthogonality condition for the vectors $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ we have

$$0.8077\ x_1^{(2)} + 0.7720\ x_2^{(2)} + x_3^{(2)} = 0$$

whence

$$x_3^{(2)} = -0.8077\ x_1^{(2)} - 0.7720\ x_2^{(2)} \tag{10}$$

Substituting this expression into system (8) and putting $x_2^{(2)} = 1$, we obtain

$$\left.\begin{aligned} x_1^{(2)} &= \frac{1}{\lambda_2}\ (2.3846\ x_1^{(2)} + 0.4560), \\ \lambda_2 &= 1.1923\ x_1^{(2)} + 4.2280 \end{aligned}\right\} \tag{11}$$

We solve system (11) by the iteration method, setting

$$x_1^{(2,\,0)} = 1 \text{ and } \lambda_2^{(0)} = 5.42$$

The results of the computations are arranged in Table 30.

TABLE 30

USING THE METHOD OF ITERATION TO COMPUTE THE EIGENVALUES AND EIGENVECTORS OF A MATRIX WHICH CORRESPOND TO THE SECOND ROOT OF THE CHARACTERISTIC EQUATION

k	$x_1^{(2k)}$	$x_2^{(2k)}$	$\lambda_2^{(k)}$	k	$x_1^{(2k)}$	$x_2^{(2k)}$	$\lambda_2^{(k)}$
0	1	1	5.42	6	0.223	1	4.494
1	0.52	1	4.85	7	0.220	1	4.490
2	0.35	1	4.64	8	0.218	1	4.488
3	0.28	1	4.56	9	0.2174	1	4.487
4	0.25	1	4.53	10	0.2171	1	4.4868
5	0.23	1	4.500	11	0.2170	1	4.4867

We can take $\lambda_2 = 4.4867$ and $x_1^{(2)} = 0.2170$, $x_2^{(2)} = 1$.

The third coordinate is determined from the orthogonal relations (10):

$$x_3^{(2)} = -0.9473$$

and so

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 0.2170 \\ 1 \\ -0.9473 \end{bmatrix}$$

The third eigenvector $\boldsymbol{x}^{(3)}$ is determined directly from the two orthogonal relations

$$\left.\begin{array}{l} 0.8077\ x_1^{(3)}+0.7720\ x_2^{(3)}+x_3^{(3)}=0, \\ 0.2170\ x_1^{(3)}+x_2^{(3)}-0.9473\ x_3^{(3)}=0 \end{array}\right\}$$

Putting $x_1^{(3)}=1$, we get $x_2^{(3)}=-0.5673$, $x_3^{(3)}=-0.3698$. Hence

$$\boldsymbol{x}^{(3)}=\begin{bmatrix} 1 \\ -0.5673 \\ -0.3698 \end{bmatrix}$$

From the last equation of system (8) we also find, for $j=3$,

$$\lambda_3=2.1260$$

For a check, form the trace of matrix A:

$$\begin{aligned} \operatorname{tr} A &= \lambda_1+\lambda_2+\lambda_3=8.3874+4.4867+2.1260= \\ &=15.0001 \approx 4+5+6 \end{aligned}$$

Note that as a rule the roots obtained by the iteration process are arranged in descending order of their moduli. The eigenvectors of the matrix are determined to within the proportionality factors, and so all the solutions of (8) are:

λ_j	$x_1^{(j)}$	$x_2^{(j)}$	$x_3^{(j)}$
8.3874	$0.8077c_1$	$0.7720c_1$	c_1
4.4867	$0.2170c_2$	c_2	$-0.9473c_2$
2.1260	c_3	$-0.5673c_3$	$-0.3698c_3$

where c_1, c_2, c_3 are arbitrary constants different from zero.

12.16 USING THE COEFFICIENTS OF THE CHARACTERISTIC POLYNOMIAL OF A MATRIX FOR MATRIX INVERSION

In Secs. 12.3 to 12.9 we gave techniques for polynomial expansion of the secular determinant of a matrix. It is comparatively simple to obtain the inverse matrix A^{-1} with the aid of the coefficients of this characteristic polynomial and by forming the powers A, A^2, ..., A^{n-1} of a nonsingular matrix A of order n. In this respect, the Leverrier method (Sec. 12.8) is particularly advantageous.

Suppose we have a nonsingular matrix A of order n. Consider its characteristic polynomial

$$\det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_{n-1}\lambda + p_n$$

According to the Cayley-Hamilton theorem (Sec. 11.2) we have

$$A^n + p_1A^{n-1} + \ldots + p_{n-1}A + p_nE = 0 \qquad (1)$$

Postmultiplying the matrix equation (1) by A^{-1}, we get

$$A^{n-1} + p_1A^{n-2} + \ldots + p_{n-1}E + p_nA^{-1} = 0 \qquad (2)$$

whence, for $p_n \neq 0$, we have

$$A^{-1} = -\frac{1}{p_n}(A^{n-1} + p_1A^{n-2} + \ldots + p_{n-1}E) \qquad (3)$$

Thus, if the coefficients of the characteristic polynomial of matrix A are known and the powers of this matrix are formed up to the $(n-1)$th inclusive, then the inverse A^{-1} can easily be computed by formula (3).

Note that if $p_n = 0$ and $p_{n-1} \neq 0$, then in order to obtain a formula containing A^{-1} it is necessary to postmultiply the matrix equation (1) by A^{-2}, etc.

Example. Find the inverse A^{-1} of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

(see Sec. 12.8, Example).

Solution. We take advantage of the earlier found powers of matrix A (Sec. 12.8):

$$A^2 = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix}$$

and

$$A^3 = \begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix}$$

Since the characteristic polynomial of matrix A is of the form

$$\det(\lambda A - E) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20$$

then from formula (3) we get

$$A^{-1} = -\frac{1}{-20}\left\{\begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix} - \right.$$

$$\left. -4\begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix} - 40\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} - 56\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right\} =$$

$$= \frac{1}{10}\left\{\begin{bmatrix} 104 & 89 & 96 & 121 \\ 89 & 74 & 77 & 96 \\ 96 & 77 & 74 & 89 \\ 121 & 96 & 89 & 104 \end{bmatrix} - \begin{bmatrix} 60 & 44 & 36 & 40 \\ 44 & 36 & 32 & 36 \\ 36 & 32 & 36 & 44 \\ 40 & 36 & 44 & 60 \end{bmatrix} - \right.$$

$$\left. - \begin{bmatrix} 20 & 40 & 60 & 80 \\ 40 & 20 & 40 & 60 \\ 60 & 40 & 20 & 40 \\ 80 & 60 & 40 & 20 \end{bmatrix} - \begin{bmatrix} 28 & 0 & 0 & 0 \\ 0 & 28 & 0 & 0 \\ 0 & 0 & 28 & 0 \\ 0 & 0 & 0 & 28 \end{bmatrix}\right\} =$$

$$= \begin{bmatrix} -0.4 & 0.5 & 0 & 0.1 \\ 0.5 & -1 & 0.5 & 0 \\ 0 & 0.5 & -1 & 0.5 \\ 0.1 & 0 & 0.5 & -0.4 \end{bmatrix}$$

As a check, we form the product

$$AA^{-1} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}\begin{bmatrix} -0.4 & 0.5 & 0 & 0.1 \\ 0.5 & -1 & 0.5 & 0 \\ 0 & 0.5 & -1 & 0.5 \\ 0.1 & 0 & 0.5 & -0.4 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = E$$

12.17 THE METHOD OF LYUSTERNIK FOR ACCELERATING THE CONVERGENCE OF THE ITERATION PROCESS IN THE SOLUTION OF A SYSTEM OF LINEAR EQUATIONS

Suppose a system of linear equations

$$Ax = b \tag{1}$$

has been reduced to a form convenient for iteration,

$$x = \beta + \alpha x \tag{1'}$$

According to the method of iteration (Sec. 8.8), the successive approximations of the solution x of system (1′) are determined from the formula

$$x^{(m)} = \beta + \alpha x^{(m-1)} \qquad (m = 1, 2, \ldots) \tag{2}$$

where $x^{(0)}$ is an arbitrary initial vector.

We assume that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the matrix α are distinct, and

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n| \tag{3}$$

The process of iteration (2) converges if

$$|\lambda_1| < 1$$

The first eigenvalue λ_1 may be approximated with the aid of methods indicated above (Secs. 12.11 and 12.12). L. A. Lyusternik [6] demonstrated that by using the eigenvalue λ_1 it is possible to substantially improve the convergence of the iteration process (2) for solving system (1′). We will now show how this is done.

For m sufficiently large we can put, approximately,

$$x \approx x^{(m)}$$

Let us estimate the error $x - x^{(m)}$. Provided the process (2) is convergent, we have

$$x = \lim_{m \to \infty} x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)})$$

and besides

$$x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)})$$

Therefore

$$\begin{aligned} x - x^{(m)} &= \sum_{k=m+1}^{\infty} (x^{(k)} - x^{(k-1)}) = \\ &= [x^{(m+1)} - x^{(m)}] + [x^{(m+2)} - x^{(m+1)}] + \ldots \end{aligned} \tag{4}$$

Since

$$\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}=[\beta+\alpha\boldsymbol{x}^{(k-1)}]-[\beta+\alpha\boldsymbol{x}^{(k-2)}]=$$
$$=\alpha(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k-2)})=\alpha^{k-1}(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}) \text{ for } k=1,\ 2,\ \ldots$$

it follows that

$$\boldsymbol{x}-\boldsymbol{x}^{(m)}=\alpha^{m}(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)})+\alpha^{m+1}(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)})+\ldots \tag{5}$$

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ be the eigenvectors of the matrix α that correspond to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and form a basis in the space E_n. Expanding the vector $\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}$ into the vectors of this basis, we get

$$\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}=c_1\boldsymbol{y}_1+c_2\boldsymbol{y}_2+\ldots+c_n\boldsymbol{y}_n$$

where c_j $(j=1, 2, \ldots, n)$ are certain definite scalars. From this,

$$\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}=\alpha^{k-1}(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)})=$$
$$=c_1\lambda^{k-1}\boldsymbol{y}_1+c_2y_2\lambda_2^{k-1}\boldsymbol{y}_2+\ldots+c_n\lambda_n^{k-1}\boldsymbol{y}_n \tag{6}$$
$$(k=m+1,\ m+2,\ \ldots)$$

Hence, we find, on the basis of (5),

$$\boldsymbol{x}-\boldsymbol{x}^{m}=c_1\lambda_1^m(1+\lambda_1+\lambda_1^2+\ldots)\boldsymbol{y}_1+$$
$$+c_2\lambda_2^m(1+\lambda_2+\lambda_2^2+\ldots)\boldsymbol{y}_2+\ldots$$
$$\ldots+c_n\lambda_n^m(1+\lambda_n+\lambda_n^2+\ldots)\boldsymbol{y}_n=$$
$$=\frac{c_1\lambda_1^m}{1-\lambda_1}\boldsymbol{y}_1+\frac{c_2\lambda_2^m}{1-\lambda_2}\boldsymbol{y}_2+\ldots+\frac{c_n\lambda_n^m}{1-\lambda_n}\boldsymbol{y}_n$$

Then, taking into account inequality (3), we obtain,

$$\boldsymbol{x}-\boldsymbol{x}^{(m)}=\frac{c_1\lambda_1^m}{1-\lambda_1}\boldsymbol{y}_1+O(\lambda_2^m) \tag{7}$$

Besides, from formula (6) we derive, for $k=m+1$,

$$\boldsymbol{x}^{(m+1)}-\boldsymbol{x}^{(m)}=c_1\lambda_1^m\boldsymbol{y}_1+O(\lambda_2^m) \tag{8}$$

and so

$$\boldsymbol{x}-\boldsymbol{x}^{(m)}=\frac{\boldsymbol{x}^{(m+1)}-\boldsymbol{x}^{(m)}}{1-\lambda_1}+O(\lambda_2^m)$$

Thus, we finally have

$$\boldsymbol{x}\approx\boldsymbol{x}^{(m)}+\frac{\boldsymbol{x}^{(m+1)}-\boldsymbol{x}^{(m)}}{1-\lambda_1} \tag{9}$$

The additional term $\frac{\boldsymbol{x}^{(m+1)}-\boldsymbol{x}^{m}}{1-\lambda_1}$ perceptibly accelerates the convergence of the iteration process (2).

Since it follows from (8) that

$$\boldsymbol{x}^{(m+1)}-\boldsymbol{x}^{(m)}=\lambda_1(\boldsymbol{x}^{(m)}-\boldsymbol{x}^{(m-1)})+O(\lambda_2^m) \tag{10}$$

formula (9) may be replaced by the following one:

$$\boldsymbol{x} \approx \boldsymbol{x}^{(m)} + \frac{\lambda_1}{1-\lambda_1}\left(\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)}\right) \tag{11}$$

Formula (11) makes it unnecessary to compute the next, in order, approximation.

On the basis of (10), the largest eigenvalue λ_1 may be determined from the formula

$$\lambda_1 \approx \frac{(\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)})_i}{(\boldsymbol{x}^{(m-1)} - \boldsymbol{x}^{(m-2)})_i} \qquad (i = 1,\ 2,\ \ldots,\ n)$$

In the case of a symmetric matrix α, we get a more exact formula by using the method of scalar products:

$$\lambda_1 \approx \frac{(\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)},\ \boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)})}{(\boldsymbol{x}^{(m-1)} - \boldsymbol{x}^{(m-2)},\ \boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)})}$$

In particular, if

$$\boldsymbol{x}^{(0)} = \boldsymbol{\beta}$$

then

$$\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)} = \alpha^{m-1}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) = \alpha^m \boldsymbol{\beta}$$

and

$$\boldsymbol{x}^{(m)} = \boldsymbol{x}^{(0)} + \sum_{k=1}^{m} \alpha^{k-1}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}) = \sum_{k=0}^{m} \alpha^k \boldsymbol{\beta}$$

Therefore

$$\lambda_1 \approx \frac{(\alpha^m \boldsymbol{\beta})_i}{(\alpha^{m-1}\boldsymbol{\beta})_i} \qquad (i = 1,\ 2,\ \ldots,\ n) \tag{12}$$

where $(\alpha^m\boldsymbol{\beta})_i$ and $(\alpha^{m-1}\boldsymbol{\beta})_i$ are the ith coordinates of the vectors $\alpha^m\boldsymbol{\beta}$ and $\alpha^{m-1}\boldsymbol{\beta}$, respectively. Similarly, if matrix α is symmetric, then

$$\lambda_1 \approx \frac{(\alpha^m\boldsymbol{\beta},\ \alpha^m\boldsymbol{\beta})}{(\alpha^{m-1}\boldsymbol{\beta},\ \alpha^m\boldsymbol{\beta})} \tag{13}$$

Example. Using the method of iteration, solve the system of equations [1]

$$\left.\begin{aligned} 0.78x_1 - 0.02x_2 - 0.12x_3 - 0.14x_4 &= 0.76, \\ -0.02x_1 + 0.86x_2 - 0.04x_3 + 0.006x_4 &= 0.08, \\ -0.12x_1 - 0.04x_2 + 0.72x_3 - 0.08x_4 &= 1.12, \\ -0.14x_1 + 0.06x_2 - 0.08x_3 + 0.74x_4 &= 0.68 \end{aligned}\right\}$$

and apply the Lyusternik method to improve the roots.

Solution. Reduce the system to a form convenient for application of the iteration method:

$$\left.\begin{aligned} x_1 &= 0.22x_1 + 0.02x_2 + 0.12x_3 + 0.14x_4 + 0.76, \\ x_2 &= 0.02x_1 + 0.14x_2 + 0.04x_3 - 0.06x_4 + 0.08, \\ x_3 &= 0.12x_1 + 0.04x_2 + 0.28x_3 + 0.08x_4 + 1.12, \\ x_4 &= 0.14x_1 - 0.06x_2 + 0.08x_3 + 0.26x_4 + 0.68 \end{aligned}\right\} \tag{14}$$

or, in matrix form,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{bmatrix} + \begin{bmatrix} 0.22 & 0.02 & 0.12 & 0.14 \\ 0.02 & 0.14 & 0.04 & -0.06 \\ 0.12 & 0.04 & 0.28 & 0.08 \\ 0.14 & -0.06 & 0.08 & 0.26 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \tag{14'}$$

whence

$$\alpha = \begin{bmatrix} 0.22 & 0.02 & 0.12 & 0.14 \\ 0.02 & 0.14 & 0.04 & -0.06 \\ 0.12 & 0.04 & 0.28 & 0.08 \\ 0.14 & -0.06 & 0.08 & 0.26 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{bmatrix}$$

Since

$$\|\alpha\|_m = \max(0.50,\ 0.26,\ 0.52,\ 0.54) = 0.54 < 1$$

the process of iteration for system (14) converges.

Using the vector $\boldsymbol{\beta}$ as the initial vector $\boldsymbol{x}^{(0)}$, we obtain, for the mth approximation $\boldsymbol{x}^{(m)}$ of the desired solution

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

the following expression

$$\boldsymbol{x}^{(m)} = \sum_{k=0}^{m} \alpha^k \boldsymbol{\beta} \tag{15}$$

Thus, in order to compute $\boldsymbol{x}^{(m)}$ we have to form successive iterations of the vector $\boldsymbol{\beta}$ by means of matrix α. We have

$$\alpha\boldsymbol{\beta} = \begin{bmatrix} 0.22 & 0.02 & 0.12 & 0.14 \\ 0.02 & 0.14 & 0.04 & -0.06 \\ 0.12 & 0.04 & 0.28 & 0.08 \\ 0.14 & -0.06 & 0.08 & 0.26 \end{bmatrix} \begin{bmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{bmatrix} = \begin{bmatrix} 0.3984 \\ 0.0304 \\ 0.4624 \\ 0.3680 \end{bmatrix},$$

$$\alpha^2\beta=\alpha\cdot\alpha\beta=\begin{bmatrix}0.22 & 0.02 & 0.12 & 0.14\\ 0.02 & 0.14 & 0.04 & -0.06\\ 0.12 & 0.04 & 0.28 & 0.08\\ 0.14 & -0.06 & 0.08 & 0.26\end{bmatrix}\begin{bmatrix}0.3984\\ 0.0304\\ 0.4624\\ 0.3680\end{bmatrix}=\begin{bmatrix}0.195264\\ 0.008640\\ 0.207936\\ 0.186624\end{bmatrix}$$

and so on.

The results of the appropriate computations are listed in Table 31.

TABLE 31

SUCCESSIVE ITERATIONS OF THE VECTOR β BY THE MATRIX α

β	$\alpha\beta$	$\alpha^2\beta$	$\alpha^3\beta$	$\alpha^4\beta$
0.76 0.08 1.12 0.68	0.3984 0.0304 0.4624 0.3680	0.195264 0.008640 0.207936 0.186624	0.09421056 0.00223488 0.09692928 0.09197568	0.04527913 0.00055572 0.04589292 0.04472340

$\alpha^5\beta$	$\alpha^6\beta$	$\alpha^7\beta$	$\alpha^8\beta$	$x^{(8)}=\sum_{k=0}^{8}\alpha^k\beta$
0.02174095 0.00013570 0.02188361 0.02160525	0.01043649 0.00003285 0.01047017 0.01040364	0.00500961 0.00000792 0.00501763 0.00500170	0.00240463 0.00000190 0.00240654 0.00240272	1.532746 0.122009 1.972937 1.410737

In formula (11) we take $m=8$. Since the matrix α is symmetric, to compute its first eigenvalue λ_1 we use the method of scalar products. We have

$$\lambda_1\approx\frac{(\alpha^8\beta,\ \alpha^8\beta)}{(\alpha^7\beta,\ \alpha^8\beta)}=$$

$$=\frac{240{,}463^2+190^2+240{,}654^2+240{,}272^2}{500{,}961\cdot240{,}463+792\cdot190+501{,}763\cdot240{,}654+500{,}170\cdot240{,}272}=$$

$$=0.480000$$

whence, taking into account that $x^{(8)}-x^{(7)}=\alpha^8\beta$, we find

$$x\approx x^8+\lambda_1\cdot\frac{\alpha^8\beta}{1-\lambda_1}=$$

$$=\begin{bmatrix}1.532746\\ 0.122009\\ 1.972937\\ 1.410737\end{bmatrix}+\frac{12}{13}\begin{bmatrix}0.002405\\ 0.000002\\ 0.002406\\ 0.002403\end{bmatrix}=\begin{bmatrix}1.534965\\ 0.122011\\ 1.975159\\ 1.412955\end{bmatrix}$$

To compare, we give the values of the roots of system (11) obtained by the Gaussian method [1]:

$$x_1 = 1.534965, \qquad x_2 = 0.122010,$$
$$x_3 = 1.975166, \qquad x_4 = 1.412955$$

Thus, whereas $\boldsymbol{x}^{(8)}$ yielded values of the roots x_i $(i = 1, 2, 3, 4)$ with a rough accuracy of $1 \cdot 10^{-3}$-$2 \cdot 10^{-3}$, the corrections of Lyusternik yield these roots with an approximate accuracy of 10^{-6}.

The Lyusternik method of accelerating convergence can also be applied to the Seidel process. As we know, the Seidel process for system (2) is a process of iteration for the equivalent system

$$\boldsymbol{x} = \boldsymbol{\beta}_1 + \alpha_1 \boldsymbol{x}$$

where the matrix α_1 is uniquely defined in terms of the matrix α (see Sec. 11.3); namely, if

$$\alpha = B + C$$

where B is a lower triangular matrix with zero diagonal and C is an upper triangular matrix, then

$$\alpha_1 = (E - B)^{-1} C$$

For this reason, if $\xi^{(m)}$ $(m = 1, 2, \ldots)$ are successive Seidel approximations of the root $\boldsymbol{x}$ of the system (2), then we can put

$$\boldsymbol{x} \approx \xi^{(m)} + \frac{\xi^{(m+1)} - \xi^{(m)}}{1 - \mu_1}$$

where μ_1 is the numerically largest eigenvalue of the matrix α_1.

There are also other methods for accelerating the convergence of iteration processes in the solution of systems of linear equations, such as the method of M. K. Gavurin [7], [8] and the method of A. A. Abramov [9].

REFERENCES FOR CHAPTER 12

[1] *V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1950, Chapter III (in Russian).
[2] *I. M. Gelfand, Lectures on Linear Algebra*, 1951, Appendix I (in Russian).
[3] *A. G. Kurosh, Course of Higher Algebra*, 1972, Chapter 6 (translated from the Russian).
[4] Harold Wayland, Expansion of Determinantal Equations into Polynomial Form, *Quarterly of Applied Math.* 1944, Vol. II, No. 4.
[5] *W. E. Milne, Numerical Calculus*, 1949, Chapter II
[6] *L. A. Lyusternik, Transactions of the Steklov Institute of Mathematics*, **20** (1947) (in Russian).
[7] M. K. Gavurin, Application of Polynomials of Best Approximation to the Acceleration of Convergence of Iterative Processes, *Uspekhi Matem. Nauk*, 5:3 (**37**) (1950), (in Russian).
[8] *I. S. Berezin and N. P. Zhidkov, Computational Methods*, 1959, Vol. 2, Chapter VIII (in Russian).
[9] *D. K. Faddeyev and V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1960, Chapter IX (in Russian).

Chapter 13

APPROXIMATE SOLUTION OF SYSTEMS OF NONLINEAR EQUATIONS

13.1 NEWTON'S METHOD

We consider, generally speaking, a nonlinear system of equations

$$\left.\begin{array}{l} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{array}\right\} \tag{1}$$

with real left members.

We write system (1) more compactly by regarding the set of arguments $x_1, x_2, \ldots, x_n$ as an n-dimensional vector:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Similarly, the set of functions $f_1, f_2, \ldots, f_n$ is also an n-dimensional vector (*vector function*):

$$\boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

The system (1) can therefore be written briefly as

$$\boldsymbol{f}(\boldsymbol{x}) = 0 \tag{1'}$$

We solve (1') by the method of successive approximations. Suppose we have found the pth approximation

$$\boldsymbol{x}^{(p)} = (x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})$$

of one of the isolated roots $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of the vector equation (1′). Then the exact root of (1′) can be represented as

$$\boldsymbol{x} = \boldsymbol{x}^{(p)} + \boldsymbol{\varepsilon}^{(p)} \tag{2}$$

where $\boldsymbol{\varepsilon}^{(p)} = (\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, \ldots, \varepsilon_n^{(p)})$ is the correction (**error of the root**).

Putting (2) into (1′), we have

$$\boldsymbol{f}(\boldsymbol{x}^{(p)} + \boldsymbol{\varepsilon}^{(p)}) = 0 \tag{3}$$

On the assumption that the function $\boldsymbol{f}(\boldsymbol{x})$ is continuously differentiable in some convex domain containing $\boldsymbol{x}$ and $\boldsymbol{x}^{(p)}$, we expand the left-hand member of equation (3) in powers of the small vector $\boldsymbol{\varepsilon}^{(p)}$ confining ourselves to linear terms,

$$\boldsymbol{f}(\boldsymbol{x}^{(p)} + \boldsymbol{\varepsilon}^{(p)}) = \boldsymbol{f}(\boldsymbol{x}^{(p)}) + \boldsymbol{f}'(\boldsymbol{x}^{(p)})\,\boldsymbol{\varepsilon}^{(p)} = 0 \tag{4}$$

or, in expanded form,

$$\left.\begin{aligned}
&f_1(x_1^{(p)}+\varepsilon_1^{(p)}, x_2^{(p)}+\varepsilon_2^{(p)}, \ldots, x_n^{(p)}+\varepsilon_n^{(p)}) =\\
&\quad = f_1(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{1x_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} +\\
&\qquad + f'_{1x_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots\\
&\qquad\qquad \ldots + f'_{1x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0,\\
&f_2(x_1^{(p)}+\varepsilon_1^{(p)}, x_2^{(p)}+\varepsilon_2^{(p)}, \ldots, x_n^{(p)}+\varepsilon_n^{(p)}) =\\
&\quad = f_2(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{2x_1}(x_1^{(p)}, x_2^{(p)} \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} +\\
&\qquad + f'_{2x_2}, (x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots\\
&\qquad\qquad \ldots + f'_{2x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0,\\
&\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\\
&f_n(x_1^{(p)}+\varepsilon_1^{(p)}, x_2^{(p)}+\varepsilon_2^{(p)}, \ldots, x_n^{(p)}+\varepsilon_n^{(p)}) =\\
&\quad = f_n(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{nx_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} +\\
&\qquad + f'_{nx_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots\\
&\qquad\qquad \ldots + f'_{nx_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}), \varepsilon_n^{(p)} = 0
\end{aligned}\right\} \tag{4′}$$

From formulas (4) and (4′) it follows that the derivative $\boldsymbol{f}'(\boldsymbol{x})$ is to be understood as the *Jacobi matrix* of the set of functions $f_1, f_2, \ldots, f_n$ with respect to the variables $x_1, x_2, \ldots, x_n$; that is,

$$\boldsymbol{f}'(\boldsymbol{x}) = W(\boldsymbol{x}) = \begin{bmatrix}
\dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \ldots & \dfrac{\partial f_1}{\partial x_n}\\
\dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \ldots & \dfrac{\partial f_2}{\partial x_n}\\
\cdot & \cdot & \cdot & \cdot\\
\dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \ldots & \dfrac{\partial f_n}{\partial x_n}
\end{bmatrix}$$

or, briefly,

$$f'(x) = W(x) = \left[\frac{\partial f_i}{\partial x_j}\right] \qquad (i, j = 1, 2, \ldots, n)$$

The system (4′) is a linear system in the corrections $\varepsilon_i^{(p)}$ $(i = 1, 2, \ldots, n)$ with matrix $W(x)$, and so formula (4) may be written as follows:

$$f(x^{(p)}) + W(x^{(p)})\,\varepsilon^{(p)} = 0$$

whence, assuming that the matrix $W(x^{(p)})$ is nonsingular, we get

$$\varepsilon^{(p)} = -W^{-1}(x^{(p)})\,f(x^{(p)})$$

Hence

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)})\,f(x^{(p)}) \qquad (p = 0, 1, 2, \ldots) \tag{5}$$

(*Newton's method*).

For the zeroth approximation $x^{(0)}$ we can take a rough value of the desired root.

Example 1. Approximate the positive solutions of the following system of equations (cf. Sec. 4.9):

$$\left.\begin{aligned} f_1(x_1, x_2) &\equiv x_1 + 3\log_{10} x_1 - x_2^2 = 0, \\ f_2(x_1, x_2) &\equiv 2x_1^2 - x_1x_2 - 5x_1 + 1 = 0, \end{aligned}\right\} \tag{6}$$

Solution. The curves defined by system (6) intersect approximately in the points $M_1(1.4, -1.5)$ and $M_2(3.4, 2.2)$. Starting with the initial approximation

$$x^{(0)} = \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix}$$

we compute the second approximations of the roots, carrying the computations to four decimal places. Setting

$$f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}$$

we have

$$f(x^{(0)}) = \begin{bmatrix} 3.4 + 3\log_{10} 3.4 - 2.2^2 \\ 2\cdot 3.4^2 - 3.4\cdot 2.2 - 5\cdot 3.4 + 1 \end{bmatrix} = \begin{bmatrix} 0.1544 \\ -0.3600 \end{bmatrix}$$

Now form the Jacobi matrix

$$W(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 + \frac{3M}{x_1} & -2x_2 \\ 4x_1 - x_2 - 5 & -x_1 \end{bmatrix}$$

where $M = 0.43429$, whence

$$W(\boldsymbol{x}^{(0)}) = \begin{bmatrix} 1+\frac{3\cdot 0.43429}{3.4} & -2\cdot 2.2 \\ 4\cdot 3.4-2.2-5 & -3.4 \end{bmatrix} = \begin{bmatrix} 1.3832 & -4.4 \\ 6.4 & -3.4 \end{bmatrix}$$

and

$$\Delta = \det W(\boldsymbol{x}^{(0)}) = 23.4571$$

Thus, the matrix $W(\boldsymbol{x}^{(0)})$ is nonsingular. Form the inverse

$$W^{-1}(\boldsymbol{x}^{(0)}) = \frac{1}{\Delta}\begin{bmatrix} -3.4 & 4.4 \\ -6.4 & 1.3832 \end{bmatrix}$$

Using formula (5), we get

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix} - \frac{1}{23.4571}\begin{bmatrix} -3.4 & 4.4 \\ -6.4 & 1.3832 \end{bmatrix}\begin{bmatrix} 0.1544 \\ -0.3600 \end{bmatrix} =$$

$$= \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix} - \frac{1}{23,4571}\begin{bmatrix} -2.10896 \\ -1.48604 \end{bmatrix} = \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix} + \begin{bmatrix} 0.0899 \\ 0.0633 \end{bmatrix} = \begin{bmatrix} 3.4899 \\ 2.2633 \end{bmatrix}$$

The subsequent approximations are found analogously. The results of the computations are listed in Table 32.

TABLE 32

SUCCESSIVE APPROXIMATIONS OF THE ROOTS OF SYSTEM (6)

i	x_1	$\varepsilon_1=\Delta x_1$	x_2	$\varepsilon_2=\Delta x_2$
0	3.4	0.0899	2.2	0.0633
1	3.4899	—0.0008	2.2633	—0.0012
2	3.4891	—0.0016	2.2621	—0.0005
3	3.4875		2.2616	

Stopping with the approximation $\boldsymbol{x}^{(3)}$, we have

$$x_1 = 3.4875, \quad x_2 = 2.2616$$

and

$$\boldsymbol{f}(\boldsymbol{x}^{(3)}) = \begin{bmatrix} 0.0002 \\ 0.0000 \end{bmatrix}$$

Example 2. Use the Newton method to approximate positive solution of the system of equations

$$\left.\begin{aligned} x^2+y^2+z^2 &= 1, \\ 2x^2+y^2-4z &= 0, \\ 3x^2-4y+z^2 &= 0 \end{aligned}\right\}$$

starting with the initial approximation

$$x_0 = y_0 = z_0 = 0.5$$

Solution. We have

$$f(\boldsymbol{x}) = \begin{bmatrix} x^2 + y^2 + z^2 - 1 \\ 2x^2 + y^2 - 4z \\ 3x^2 - 4y + z^2 \end{bmatrix}$$

whence

$$f(\boldsymbol{x}^{(0)}) = \begin{bmatrix} 0.25 + 0.25 + 0.25 - 1 \\ 0.50 + 0.25 - 2.00 \\ 0.75 - 2.00 + 0.25 \end{bmatrix} = \begin{bmatrix} -0.25 \\ -1.25 \\ -1.00 \end{bmatrix}$$

Form the Jacobi matrix

$$W(\boldsymbol{x}) = \begin{bmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{bmatrix}$$

We have

$$W(\boldsymbol{x}^{(0)}) = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{bmatrix}$$

and

$$\det W(\boldsymbol{x}^{(0)}) = \begin{vmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{vmatrix} = -40$$

The inverse matrix is

$$W^{-1}(\boldsymbol{x}^{(0)}) = -\frac{1}{40}\begin{bmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{bmatrix} = \begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & -\frac{3}{20} \\ \frac{11}{40} & -\frac{7}{40} & \frac{1}{40} \end{bmatrix}$$

Using formula (5), we obtain the first approximation:

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - W^{-1}(\boldsymbol{x}^{(0)}) f(\boldsymbol{x}^{(0)}) =$$

$$= \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & \frac{-3}{20} \\ \frac{11}{40} & \frac{-7}{40} & \frac{1}{40} \end{bmatrix} \begin{bmatrix} -0.25 \\ -1.25 \\ -1.00 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.375 \\ 0 \\ -0.125 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 0.500 \\ 0.375 \end{bmatrix}$$

Then we compute the second approximation $\boldsymbol{x}^{(2)}$ to get

$$\boldsymbol{f}(\boldsymbol{x}^{(1)})=\begin{bmatrix}0.875^2+0.500^2+0.375^2-1\\ 2\cdot0.875^2+0.500^2-4\cdot0.375\\ 3\cdot0.875^2-4\cdot0.500+0.375^2\end{bmatrix}=\begin{bmatrix}0.15625\\ 0.28125\\ 0.43750\end{bmatrix}$$

and

$$W(\boldsymbol{x}^{(1)})=\begin{bmatrix}2\cdot0.875 & 2\cdot0.500 & 2\cdot0.375\\ 4\cdot0.875 & 2\cdot0.500 & -4\\ 6\cdot0.875 & -4 & 2\cdot0.375\end{bmatrix}=$$

$$=\begin{bmatrix}1.750 & 1 & 0.750\\ 3.500 & 1 & -4\\ 5.250 & -4 & 0.750\end{bmatrix}$$

whence

$$\det W(\boldsymbol{x}^{(1)})=\begin{vmatrix}1.750 & 1 & 0.750\\ 3.500 & 1 & -4\\ 5.250 & -4 & 0.750\end{vmatrix}=$$

$$=\begin{vmatrix}1.750 & 1 & 0.750\\ 1.750 & 0 & -4.750\\ 12.250 & 0 & 3.750\end{vmatrix}=-64.75$$

and

$$W^{-1}(\boldsymbol{x}^{(1)})=-\frac{1}{64.75}\begin{bmatrix}-15.25 & -3.75 & -4.75\\ -23.625 & -2.6250 & 9.625\\ -19.25 & 12.25 & -1.75\end{bmatrix}$$

Using formula (5), we obtain

$$\boldsymbol{x}^{(2)}=\boldsymbol{x}^{(1)}-W^{-1}(\boldsymbol{x}^{(1)})\boldsymbol{f}(\boldsymbol{x}^{(1)})=$$

$$=\begin{bmatrix}0.875\\ 0.500\\ 0.375\end{bmatrix}+\frac{1}{64.75}\begin{bmatrix}-15.25 & -3.75 & -4.75\\ -23.625 & -2.6250 & 9.625\\ -19.25 & 12.25 & -1.75\end{bmatrix}\begin{bmatrix}0.15625\\ 0.28125\\ 0.43750\end{bmatrix}=$$

$$=\begin{bmatrix}0.875\\ 0.500\\ 0.375\end{bmatrix}-\begin{bmatrix}0.08519\\ 0.00338\\ 0.00507\end{bmatrix}=\begin{bmatrix}0.78981\\ 0.49662\\ 0.36993\end{bmatrix}$$

The subsequent approximations are found similarly:

$$\boldsymbol{x}^{(3)}=\begin{bmatrix}0.78521\\ 0.49662\\ 0.36992\end{bmatrix},\quad \boldsymbol{f}(\boldsymbol{x}^{(3)})=\begin{bmatrix}0.00001\\ 0.00004\\ 0.00005\end{bmatrix}$$

and so forth.

Stopping with the third approximation, we get

$$x = 0.7852, \quad y = 0.4966, \quad z = 0.3699$$

13.2 GENERAL REMARKS ON THE CONVERGENCE OF THE NEWTON PROCESS

In Sec. 13.1 we presented a formal aspect of the Newton method. The conditions of convergence of this method for a system have been investigated by Willers, Stenin, Ostrowski, Kantorovich, and others. Below we give a special case of the Kantorovich theorem (Theorem 1) [1] on the convergence of the Newton process in functional spaces as applied to finite systems of nonlinear equations; for the sake of simplicity we use rough estimates. Following L. V. Kantorovich, we also establish the rapidity of convergence of the Newton process, the uniqueness of the root of the system and the stability of the process with respect to choice of the initial approximation (Theorems 2 to 4). As a particular case, we obtain the Ostrowski theorem [2] on the convergence of the Newton process for an equation with an analytic complex right-hand member.

In the sequel it will be convenient to regard the sets of functions as *vector functions* or *matrix functions*. To simplify the presentation, we will generalize the concept of a derivative to these cases.

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \cdot \\ \cdot \\ \cdot \\ f_n(\boldsymbol{x}) \end{bmatrix}$$

where $f_i \in C^{(1)}$ $(i = 1, 2, \ldots, n)$.

Definition 1. The derivative $\boldsymbol{f}'(\boldsymbol{x})$ is understood to mean the Jacobi matrix of the set of functions f_i $(i = 1, \ldots, n)$ with respect to the variables $x_1, \ldots, x_n$, that is,

$$\boldsymbol{f}'(\boldsymbol{x}) = \left[\frac{\partial f_i}{\partial x_j}\right] \tag{1}$$

The matrix function

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} f_{11}(\boldsymbol{x}) & \ldots & f_{1r}(\boldsymbol{x}) \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ f_{n1}(\boldsymbol{x}) & \ldots & f_{nr}(\boldsymbol{x}) \end{bmatrix}$$

may be regarded as a set of m vector functions

$$\boldsymbol{F}_1(\boldsymbol{x})=\begin{bmatrix} f_{11}(\boldsymbol{x}) \\ \cdot \\ \cdot \\ \cdot \\ f_{n1}(\boldsymbol{x}) \end{bmatrix}, \ \ldots, \ \boldsymbol{F}_r(\boldsymbol{x})=\begin{bmatrix} f_{1r}(\boldsymbol{x}) \\ \cdot \\ \cdot \\ \cdot \\ f_{nr}(\boldsymbol{x}) \end{bmatrix}$$

Therefore, it is natural to take the derivative $\boldsymbol{F}'(\boldsymbol{x})$ as meaning the set

$$\boldsymbol{F}'(\boldsymbol{x})=[\boldsymbol{F}_1'(\boldsymbol{x}) \ \ldots \ \boldsymbol{F}_r'(\boldsymbol{x})]$$

where

$$\boldsymbol{F}_k'(\boldsymbol{x})=\begin{bmatrix} \frac{\partial f_{1k}}{\partial x_1} & \ldots & \frac{\partial f_{1k}}{\partial x_n} \\ \cdot & \cdot \ \cdot \ \cdot \ \cdot & \cdot \\ \frac{\partial f_{nk}}{\partial x_1} & \ldots & \frac{\partial f_{nk}}{\partial x_n} \end{bmatrix}$$

are Jacobi matrices ($k=1, 2, \ldots, r$).

Definition 2. If $\boldsymbol{F}(\boldsymbol{x})=[f_{ij}(\boldsymbol{x})]$ is a functional matrix of dimensions $\boldsymbol{n}\times\boldsymbol{r}$ and $f_{ij}(\boldsymbol{x})\in C^{(1)}$, then

$$\boldsymbol{F}'(\boldsymbol{x})=[\boldsymbol{F}_k'(\boldsymbol{x})] \tag{2}$$

where

$$\boldsymbol{F}_k'(\boldsymbol{x})=\left[\frac{\partial f_{ik}}{\partial x_j}\right] \qquad (i, j=1, 2, \ldots, n; \quad k=1, 2, \ldots, r)$$

In particular, if the vector function $\boldsymbol{f}(\boldsymbol{x})=[f_i(\boldsymbol{x})]$ is such that $f_i(\boldsymbol{x})\in C^{(2)}$, then

$$\boldsymbol{f}''(\boldsymbol{x})=[W_1(\boldsymbol{x}) \ \ldots \ W_n(\boldsymbol{x})]$$

where

$$W_k(\boldsymbol{x})=\left[\frac{\partial^2 f_i}{\partial x_k \partial x_j}\right] \qquad (k=1, 2, \ldots, n)$$

In this section we use the m-norm (Sec. 7.7) for estimating matrices; the subscript m will be omitted for brevity:

$$\|\boldsymbol{f}(\boldsymbol{x})\|=\max_i |f_i(\boldsymbol{x})|,$$

$$\|\boldsymbol{f}'(\boldsymbol{x})\|=\max_i \sum_{j=1}^{n}\left|\frac{\partial f_i(x)}{\partial \boldsymbol{x}_j}\right|,$$

$$\|\boldsymbol{f}''(\boldsymbol{x})\|=\max_k \|W_k(\boldsymbol{x})\|=\max_k \left\{\max_i \sum_{j=1}^{n}\left|\frac{\partial^2 f_i(x)}{\partial x_k \partial x_j}\right|\right\}, \text{ etc.}$$

Similarly

$$\|F(x)\| = \max_{i} \sum_{j=1}^{n} |f_{ij}(x)|$$

$$\|F'(x)\| = \max_{i,\, j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(x)}{\partial x_k} \right|^{1)}$$

First, we will derive several estimates, similar to the mean-value theorem, for the m-norms of the differences of values of matrix functions, which will be useful in the sequel (cf. [1]).

Lemma 1. *If*

$$F(x) = [f_{ij}(x)] \qquad (n \times r)$$

where $f_{ij}(x)$ *are continuous, together with their first-order partial derivatives, in a convex domain containing the points* x *and* $x + \Delta x$, *then*

$$\|F(x+\Delta x) - F(x)\| \leqslant r \|\Delta x\| \cdot \|F'(\xi)\| \tag{3}$$

where $\xi = x + \theta \Delta x$, $0 < \theta < 1$ *and the matrix norm is to be understood in the sense of the* m-*norm.*

Proof. Using the Taylor formula, we obtain

$$F(x+\Delta x) - F(x) = [f_{ij}(x+\Delta x) - f_{ij}(x)] = \left[\sum_{k=1}^{n} \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \Delta x_k \right]$$

where $\xi_{ij} = x + \theta_{ij} \Delta x$, $0 < \theta_{ij} < 1$; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, r$. Whence, fixing x and $x + \Delta x$, we get

$$\|F(x+\Delta x) - F(x)\| = \max_{i} \sum_{j=1}^{r} \left| \sum_{k=1}^{n} \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \Delta x_k \right| \leqslant$$

$$\leqslant \max_{i} \sum_{j=1}^{r} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right| |\Delta x_k| \leqslant$$

$$\leqslant \max_{k} |\Delta x_k| \cdot \sum_{j=1}^{r} \max_{i,\, j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right| =$$

$$= r \|\Delta x\| \max_{i,\, j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right|.$$

---

1) Since, obviously, for any finite set of numbers $\{a_{ij}\}$ we have

$$\max_{i} (\max_{j} a_{ij}) = \max_{i,\, j} a_{ij}$$

Since the number of pairs (i, j) is finite, there is a pair (p, q) such that

$$\max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\right| = \sum_{k=1}^{n}\left|\frac{\partial f_{pq}(\xi_{pq})}{\partial x_k}\right| \leqslant \max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{pq})}{\partial x_k}\right| = \|F'(\xi)\|$$

where $\xi = \xi_{pq}$.

Thus

$$\|F(x+\Delta x) - F(x)\| \leqslant r\|\Delta x\|\,\|F'(\xi)\|$$

which completes the proof.

Corollary 1. If

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

then

$$\|f(x+\Delta x) - f(x)\| \leqslant \|\Delta x\| \cdot \|f'(\xi)\|$$

where $\xi = x + \theta\Delta x$ and $0 < \theta < 1$.

Here $r = 1$.

Corollary 2. For $f(x) \in C^{(2)}$ we have

$$\|f'(x+\Delta x) - f'(x)\| \leqslant n\|\Delta x\|\,\|f''(\xi)\|$$

where $\xi = x + \theta\Delta x$ and $0 < \theta < 1$.

Lemma 2. *If*

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} \in C^{(2)}$$

in a convex domain containing the points x *and* $x + \Delta x$, *then*

$$\|f(x+\Delta x) - f(x) - f'(x)\Delta x\| = \frac{1}{2}n\|\Delta x\|^2 \cdot \|f''(\xi)\| \tag{4}$$

where $\xi = x + \theta\Delta x$ *and* $0 < \theta < 1$.

Proof. Using the two-term Taylor formula, we obtain

$$\begin{aligned}\|f(x+\Delta x) - f(x) - f'(x)\Delta x\| &= \\ &= \|[f_i(x+\Delta x) - f_i(x) - df_i(x_i)]\| = \\ &= \frac{1}{2}\left\|\left[\sum_{j,k}\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\Delta x_j\,\Delta x_k\right]\right\| \leqslant \\ &\leqslant \frac{1}{2}\left\|\left[\sum_{j}|\Delta x_j|\sum_{k}\left|\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\right||\Delta x_k|\right]\right\| \leqslant\end{aligned}$$

$$\leqslant \frac{1}{2} \max_j |\Delta x_j| \cdot \max_k |\Delta x_k| \cdot \left\| \left[\sum_j \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \right] \right\| =$$

$$= \frac{1}{2} \| \Delta \boldsymbol{x} \|^2 \left\| \left[\sum_j \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \right] \right\| \tag{5}$$

where $\xi_i = \boldsymbol{x} + \theta_i \Delta \boldsymbol{x}$, $0 < \theta_i < 1$.

Since

$$\sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,\,j} \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| =$$

$$= \sum_k \left| \frac{\partial^2 f_p(\xi_p)}{\partial x_q \partial x_k} \right| \leqslant \max_{i,\,j} \sum_k \left| \frac{\partial^2 f_i(\xi_p)}{\partial x_j \partial x_k} \right| = \| \boldsymbol{f}''(\xi_p) \|$$

then from inequality (5) we get (taking into consideration the meaning of the norm)

$$\| \boldsymbol{f}(\boldsymbol{x} + \Delta \boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x}) \Delta \boldsymbol{x} \| \leqslant \frac{1}{2} \| \Delta \boldsymbol{x} \|^2 [\| \boldsymbol{f}''(\xi) \|] =$$

$$= \frac{n}{2} \| \Delta \boldsymbol{x} \|^2 \| \boldsymbol{f}''(\xi) \|$$

where $\xi = \xi_p = \boldsymbol{x} + \theta \Delta \boldsymbol{x}$ and $0 < \theta < 1$.

*13.3 THE EXISTENCE OF ROOTS OF A SYSTEM AND THE CONVERGENCE OF THE NEWTON PROCESS

Theorem 1. *Given a nonlinear system of algebraic or transcendental equations with real coefficients:*

$$\boldsymbol{f}(\boldsymbol{x}) = 0 \tag{1}$$

where the vector function

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \cdot \;\cdot \;\cdot \;\cdot \;\cdot \;\cdot \\ f_n(x_1, \ldots, x_n) \end{bmatrix}$$

is defined and continuous, together with its partial derivatives of first and second order, in a domain ω, *that is*

$$\boldsymbol{f}(\boldsymbol{x}) \in C^{(2)}(\omega)$$

Suppose $\mathrm{x}^{(0)}$ *is a point lying in* ω *together with its closed* $\mathcal{H}$*-neighbourhood:*

$$\overline{U}_{\mathcal{H}}(\boldsymbol{x}^{(0)})=\{\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|\leqslant\mathcal{H}\}\subset\omega$$

where the norm is to be understood as the m*-norm*[1] (see Sec. 7.7), *the following conditions being valid:*

(1) *the Jacobi matrix* $\mathrm{W}(\mathrm{x})=\left[\frac{\partial \mathrm{f}_i}{\partial \mathrm{x}_j}\right]$ *has the inverse* $\Gamma_0=\mathrm{W}^{-1}(\mathrm{x}^{(0)})$ *for* $\mathrm{x}=\mathrm{x}^{(0)}$, *where*

$$\|\Gamma_0\|\leqslant A_0{}^{2)}$$

(2) $\|\Gamma_0\boldsymbol{f}(\boldsymbol{x}^{(0)})\|\leqslant B_0\leqslant\frac{\mathcal{H}}{2}$,

(3) $\sum\limits_{k=1}^{n}\left|\frac{\partial^2 f_i(\boldsymbol{x})}{\partial x_j\,\partial x_k}\right|\leqslant C$

when i, j = 1, 2, ..., n *and* $\mathrm{x}\in\overline{U}_{\mathcal{H}}(\mathrm{x}^{(0)})$,

(4) *the constants* A_0, B_0 *and* C *satisfy the inequality*

$$\mu_0=2nA_0B_0C\leqslant 1 \tag{2}$$

Then the Newton process

$$\boldsymbol{x}^{(p+1)}=\boldsymbol{x}^{(p)}-W^{-1}(\boldsymbol{x}^{(p)})\,\boldsymbol{f}(\boldsymbol{x}^{(p)}) \tag{3}$$

$(p=0,\ 1,\ 2,\ \ldots)$ *converges for the initial approximation* $\mathrm{x}^{(0)}$ *and the limiting vector*

$$\boldsymbol{x}^*=\lim_{p\to\infty}\boldsymbol{x}^{(p)}$$

is a solution of system (1) *such that*

$$\|\boldsymbol{x}^*-\boldsymbol{x}^{(0)}\|\leqslant 2B_0\leqslant\mathcal{H}$$

Proof. We introduce the notation

$$h_p=\|\boldsymbol{x}^{(p+1)}-\boldsymbol{x}^{(p)}\|=\max_k|x_k^{(p+1)}-x_k^{(p)}|,$$

$$\Gamma_p=W^{-1}(\boldsymbol{x}^{(p)})\qquad(p=0,\ 1,\ 2,\ \ldots)$$

From formula (3) we have

$$h_p=\|\Gamma_p\boldsymbol{f}(\boldsymbol{x}^{(p)})\|$$

Proceeding from the conditions (1) to (4), we obtain estimates for the quantities Γ_p and $\Gamma_p\boldsymbol{f}(\boldsymbol{x}^{(p)})$.

---

[1] That is, if $A=[a_{ij}]$, then

$$\|A\|=\|A\|_m=\max_i\sum_j|a_{ij}|$$

[2] In other words, if $W(\boldsymbol{x}^{(0)})=[a_{ij}]$, then $\Gamma_0=W^{-1}(\boldsymbol{x}^{(0)})=\left[\frac{A_{ij}}{\Delta}\right]$, where A_{ij} are the cofactors of the elements a_{ij} and $\Delta=\det[a_{ij}]$ and, consequently,

$$\|\Gamma_0\|=\max_i\frac{1}{|\Delta|}\sum_{j=1}^{n}|A_{ji}|$$

First consider the case $p=1$. Using Condition (2), we have

$$h_0 = \| x^{(1)} - x^{(0)} \| = \| W^{-1}(x^{(0)}) f(x^{(0)}) \| \leqslant B_0 \leqslant \frac{\mathcal{H}}{2}$$

and so

$$h_0 \leqslant B_0$$

and

$$\overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)}) \subset \overline{U}_{\mathcal{H}}(x^{(0)})$$

To estimate $\Gamma_1 = W^{-1}(x^{(1)})$, take advantage of the relation $(AB)^{-1} = B^{-1}A^{-1}$, and represent this quantity as

$$\Gamma_1 = [W(x^{(0)}) \cdot \Gamma_0 W(x^{(1)})]^{-1} = [\Gamma_0 W(x^{(1)})]^{-1} \cdot \Gamma_0 \tag{4}$$

Taking into account Condition (1) of the theorem, we have

$$\| E - \Gamma_0 W(x^{(1)}) \| = \| \Gamma_0 [W(x^{(0)}) - W(x^{(1)})] \| \leqslant$$
$$\leqslant \| \Gamma_0 \| \, \| W(x^{(0)}) - W(x^{(1)}) \| \leqslant A_0 \| W(x^{(1)}) - W(x^{(0)}) \|$$

Since from Condition (3) follows

$$\| f''(x) \| = \max_{i,\, j} \sum_{k=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k} \right| \leqslant C$$

then by virtue of Corollary 2 of Lemma 1 we have

$$\| W(x^{(1)}) - W(x^{(0)}) \| = \| f'(x^{(1)}) - f'(x^{(0)}) \| \leqslant$$
$$\leqslant n \| x^{(1)} - x^{(0)} \| C \leqslant n B_0 C$$

and so

$$\| E - \Gamma_0 W(x^{(1)}) \| \leqslant n A_0 B_0 C = \frac{\mu_0}{2} \leqslant \frac{1}{2}$$

Consequently (Sec. 7.10, Theorem 5, Corollary), there exists the inverse matrix

$$[\Gamma_0 W(x^{(1)})]^{-1} = \{E - (E - \Gamma_0 W(x^{(1)}))\}^{-1}$$

And since $\| E \| = \| E \|_m = 1$, it follows that

$$\| [\Gamma_0 W(x^{(1)})]^{-1} \| \leqslant \frac{1}{1 - \frac{\mu_0}{2}} \leqslant 2 \tag{5}$$

We now derive from formula (4) that

$$\| \Gamma_1 \| \leqslant \| [\Gamma_0 W(x^{(1)})]^{-1} \| \, \| \Gamma_0 \| \leqslant 2A_0 = A_1 \tag{6}$$

Formula (3) implies

$$f(x^{(0)} + f')(x^{(0)})(x^{(1)} - x^{(0)}) = 0$$

whence, on the basis of Lemma 2, we have

$$\|f(x^{(1)})\| = \|f(x^{(1)}) - f(x^{(0)}) - f'(x^{(0)})(x^{(1)} - x^{(0)})\| \leqslant$$
$$\leqslant \frac{1}{2} n \|x^{(1)} - x^{(0)}\|^2 \|f''(\xi)\| \leqslant \frac{1}{2} nB_0^2 C$$

where

$$\xi = x^{(0)} + \theta(x^{(1)} - x^{(0)}) \text{ and } 0 < \theta < 1.$$

Therefore, taking into account inequality (6), we obtain

$$\|\Gamma_1 f(x^{(1)})\| \leqslant \|\Gamma_1\| \|f(x^{(1)})\| \leqslant$$
$$\leqslant 2A_0 \cdot \frac{1}{2} nB_0^2 C = nA_0 B_0^2 C = \frac{1}{2} \mu_0 B_0 = B_1 \tag{7}$$

Thus, for point $x^{(1)}$ we have

$$\overline{U}_{\frac{\mathscr{H}}{2}}(x^{(1)}) \subset \overline{U}_{\mathscr{H}}(x^{(0)}) \subset \omega$$

and, besides,

$$\|\Gamma_1\| \leqslant A_1, \qquad h_1 = \|\Gamma_1 f(x^{(1)})\| \leqslant B_1$$

where

$$A_1 = 2A_0,$$
$$B_1 = \frac{1}{2} \mu_0 B_0 \leqslant \frac{\mathscr{H}}{4}$$

whence we obtain

$$\mu_1 = 2nA_1 B_1 C = 2n \cdot 2A_0 \cdot \frac{1}{2} \mu_0 B_0 C = \mu_0 \cdot 2nA_0 B_0 C = \mu_0^2 \leqslant 1 \tag{8}$$

Thus we are again in the conditions of the theorem with the sole difference that instead of the neighbourhood $\overline{U}_{\mathscr{H}}(x^{(0)})$ we have the neighbourhood $\overline{U}_{\frac{\mathscr{H}}{2}}(x^{(1)})$ imbedded in the former.

Repeating similar arguments, we establish that the successive approximations $x^{(p)}$ $(p = 1, 2, \ldots)$ are meaningful and such that

$$\overline{U}_{\mathscr{H}}(x^{(0)}) \supset \overline{U}_{\frac{\mathscr{H}}{2}}(x^{(1)}) \supset \ldots \supset \overline{U}_{\frac{\mathscr{H}}{2^p}}(x^{(p)}) \supset \ldots$$

Also

$$\|\Gamma_p\| = \|W^{-1}(x^{(p)})\| \leqslant A_p,$$
$$\|\Gamma_p f(x^{(p)})\| = \|x^{(p+1)} - x^{(p)}\| \leqslant B_p$$

where the constants A_p and B_p are connected by the recurrence relations

$$\left.\begin{aligned} A_p &= 2A_{p-1}, \\ B_p &= \frac{1}{2} \mu_{p-1} B_{p-1} \end{aligned}\right\} \tag{9}$$

and

$$\mu_p = 2nA_pB_pC \qquad (p=1, 2, \ldots) \tag{10}$$

We will show that the Cauchy test (Sec. 7.9) is valid for the sequence of approximations $\boldsymbol{x}^{(p)}$ $(p=0, 1, 2, \ldots)$. Indeed, for $q>0$ we have

$$\boldsymbol{x}^{(p+q)} \in \overline{U}_{\frac{\mathscr{H}}{2p}}(\boldsymbol{x}^{(p)})$$

and so

$$\| \boldsymbol{x}^{(p+q)} - \boldsymbol{x}^{(p)} \| \leqslant \frac{\mathscr{H}}{2^p} < \varepsilon$$

if $p>N$ and $q>0$, which is equivalent to the Cauchy test. From this it follows that the limit

$$\lim_{p\to\infty} \boldsymbol{x}^{(p)} = \boldsymbol{x}^* \in \overline{U}_{\mathscr{H}}(\boldsymbol{x}^{(0)})$$

exists.

Now let us assure ourselves that $\boldsymbol{x}^*$ is a solution of system (1). From the relation (3) we have

$$\boldsymbol{f}(\boldsymbol{x}^{(p)}) + W(\boldsymbol{x}^{(p)})(\boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{(p)}) = \boldsymbol{0}$$

Passing to the limit in this equation as $p\longrightarrow\infty$ and noting that, in the process,

$$\boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{(p)} \longrightarrow \boldsymbol{0}$$

and also that $W(\boldsymbol{x}^{(p)})$ is continuous and bounded in $\overline{U}_{\mathscr{H}}(\boldsymbol{x}^{(0)})$, we have

$$\lim_{p\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(p)}) = \boldsymbol{0}$$

Whence, by virtue of the continuity of the function $\boldsymbol{f}(\boldsymbol{x})$, we obtain

$$\boldsymbol{f}\left(\lim_{p\to\infty} \boldsymbol{x}^{(p)}\right) = \boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{0}$$

That is, $\boldsymbol{x}^*$ is a solution of system (1). Besides,

$$\| \boldsymbol{x}^* - \boldsymbol{x}^{(0)} \| = \left\| \sum_{p=0}^{\infty} [\boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{(p)}] \right\| \leqslant$$

$$\leqslant \sum_{p=0}^{\infty} \| \boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{(p)} \| \leqslant \sum_{p=0}^{\infty} B_p \leqslant B_0 + \frac{B_0}{2} + \ldots = 2B_0 \leqslant \mathscr{H}$$

The proof of the theorem is complete.

Note 1. If $\boldsymbol{f}(\boldsymbol{x}) \in C^{(2)}(\omega)$ and system (1) has, in the domain ω, a simple solution $\boldsymbol{x}^*$, that is, such that

$$\boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{0}, \quad \boldsymbol{f}'(\boldsymbol{x}^*) = W(\boldsymbol{x}^*) \neq 0$$

then the conditions of Theorem 1 are clearly valid for every point $\boldsymbol{x}^{(0)}$ sufficiently close to $\boldsymbol{x}^*$.

To verify Condition (2) it is useful to note that B_0 yields an estimate of the divergence between the initial and the first approximation of the Newton process:

$$\|\Gamma_0 \boldsymbol{f}(\boldsymbol{x}^{(0)})\| = \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \leqslant B_0$$

and so this inequality can readily be verified as soon as the approximation $\boldsymbol{x}^{(1)}$ is found.

Note 2. Analogous statements are obtained for the theorem of convergence if we use the norm $\|A\|_l$ or $\|A\|_k$ in place of the norm $\|A\|_m$.

*13.4 THE RAPIDITY OF CONVERGENCE OF THE NEWTON PROCESS

Theorem 2. *If Conditions* (1) *to* (4) *of Theorem* 1 (Sec. 13.3) *are fulfilled, then the following inequality holds true for the successive approximations* $\mathbf{x}^{(p)}$ $(p=0, 1, 2, \ldots)$:

$$\|\boldsymbol{x}^* - \boldsymbol{x}^{(p)}\| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p - 1} B_0$$

where $\mathbf{x}^*$ *is a solution of the system and* μ_0 *is found from formula* (2) of Sec. 13.3.

Proof. Using the relations (9) and (10) of Sec. 13.3, we have

$$\mu_p = 2nA_pB_pC = 2n \cdot 2A_{p-1} \cdot \frac{1}{2}\mu_{p-1}B_{p-1} \cdot C =$$
$$= \mu_{p-1} \cdot 2nA_{p-1}B_{p-1}C = \mu_{p-1}^2$$

From this we obtain

$$\left.\begin{aligned} &\mu_1 = \mu_0^2, \\ &\mu_2 = \mu_1^2 = \mu_0^4, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ &\mu_p = \mu_0^{2^p} \end{aligned}\right\} \tag{1}$$

Furthermore,

$$B_p = \frac{1}{2}\mu_{p-1}B_{p-1} = \frac{1}{2}\mu_0^{2^{p-1}}B_{p-1}$$

and so

$$B_p = \frac{1}{2}\mu_0^{2^{p-1}} \cdot \frac{1}{2}\mu_0^{2^{p-2}} \cdots \frac{1}{2}\mu_0^{2^0}B_0 =$$
$$= \left(\frac{1}{2}\right)^p \cdot \mu_0^{2^{p-1}+2^{p-2}+\cdots+1}B_0 = \left(\frac{1}{2}\right)^p \mu_0^{2^p-1}B_0 \tag{2}$$

Since

$$\| x^{(p+1)} - x^{(p)} \| \leqslant B_p$$

we have, for $q > 1$,

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \| x^{(p+1)} - x^{(p)} \| +$$
$$+ \| x^{(p+2)} - x^{(p+1)} \| + \dots + \| x^{(p+q)} - x^{(p+q-1)} \| \leqslant$$
$$\leqslant B_p + B_{p+1} + \dots + B_{p+q-1} =$$
$$= \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 + \left(\frac{1}{2}\right)^{p+1} \mu_0^{2^{p+1}-1} B_0 + \dots + \left(\frac{1}{2}\right)^{p+q-1} \mu_0^{2^{p+q-1}-1} B_0 =$$
$$= \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \left[1 + \frac{1}{2} \cdot \mu_0^{2^p} + \dots + \left(\frac{1}{2}\right)^{q-1} \mu_0^{2^p(2^{q-1}-1)}\right]$$

Whence, taking into account that $\mu_0 \leqslant 1$, we obtain

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \left[1 + \frac{1}{2} + \dots + \left(\frac{1}{2}\right)^{q-1}\right] \leqslant$$
$$\leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0$$

Passing to the limit as $q \to \infty$, we finally get

$$\| x^* - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0 \leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} \mathcal{H}$$

where

$$\mu_0 = 2nA_0B_0C \leqslant 1$$

Thus, for $\mu_0 < 1$ the convergence of the Newton process is superfast. For $p = 0$ we have

$$\| x^* - x^{(0)} \| \leqslant 2B_0 \leqslant \mathcal{H}$$

*13.5 UNIQUENESS OF SOLUTION

Theorem 3. *Given the Conditions* (1) *to* (4) *of Theorem* 1 (Sec. 13.3), *there is, in the domain*

$$\| x - x^{(0)} \| \leqslant 2B_0 \tag{1}$$

a unique solution of the system (1) (Sec. 13.3).

Proof. Suppose that besides the solution x^* of system (1) of Sec. 13.3, which solution is defined by the Newton process, there is another solution x^{**} of the system such that

$$\| x^{**} - x^{(0)} \| \leqslant 2B_0 \tag{2}$$

The successive approximations $x^{(p)}$ ($p = 0, 1, 2, \dots$) of the Newton process lie in the neighbourhood (1) and satisfy the condition

$$f(x^{(p)}) + W_p (x^{(p+1)} - x^{(p)}) = 0$$

where

$$W_p = W(\boldsymbol{x}^{(p)})$$

From this, taking into consideration that

$$\boldsymbol{f}(\boldsymbol{x}^{**}) = 0$$

we get

$$W_p(\boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{**}) = \boldsymbol{f}(\boldsymbol{x}^{**}) - \boldsymbol{f}(\boldsymbol{x}^{(p)}) - W_p(\boldsymbol{x}^{**} - \boldsymbol{x}^{(p)})$$

and, hence,

$$\boldsymbol{x}^{(p+1)} - \boldsymbol{x}^{**} = \Gamma_p[\boldsymbol{f}(\boldsymbol{x}^{**}) - \boldsymbol{f}(\boldsymbol{x}^{(p)}) - W_p(\boldsymbol{x}^{**} - \boldsymbol{x}^{(p)})]$$

where

$$\Gamma_p = W_p^{-1}$$

Estimating by the norm, we get

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p+1)} \| \leqslant \| \Gamma_p \| \, \| \boldsymbol{f}(\boldsymbol{x}^{**}) - \boldsymbol{f}(\boldsymbol{x}^{(p)}) - W_p(\boldsymbol{x}^{**} - \boldsymbol{x}^{(p)}) \|$$

According to the notations of Sec. 13.3 (see Theorem 1),

$$\| \Gamma_p \| \leqslant A_p$$

Applying Lemma 2 of Sec. 13.2, we get the inequality

$$\| \boldsymbol{f}(\boldsymbol{x}^{**}) - \boldsymbol{f}(\boldsymbol{x}^{(p)}) - W_p(\boldsymbol{x}^{**} - \boldsymbol{x}^{(p)}) \| \leqslant \frac{1}{2} nC \| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p)} \|^2$$

where the constant C is defined from Condition (3) of Theorem 1. For this reason

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p+1)} \| \leqslant \frac{1}{2} nA_pC \| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p)} \|^2 \qquad (p = 0, 1, 2, \ldots) \quad (3)$$

Putting $p = 0$ in inequality (3) and using inequality (2), we obtain

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(1)} \| \leqslant \frac{1}{2} nA_0C \| \boldsymbol{x}^{**} - \boldsymbol{x}^{(0)} \|^2 \leqslant 2nA_0B_0^2C$$

or, introducing the numbers defined by the relations

$$\left. \begin{aligned} \mu_p &= 2nA_pB_pC \\ B_{p+1} &= \frac{1}{2} \mu_p B_p \end{aligned} \right\} \quad (p = 0, 1, 2, \ldots) \qquad (4)$$

we find

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(1)} \| \leqslant \mu_0 B_0 = 2B_1 \qquad (5)$$

Similarly, for $p = 1$ we derive from formulas (3), (5) and (4)

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(2)} \| \leqslant \frac{1}{2} nA_1C \| \boldsymbol{x}^{**} - \boldsymbol{x}^{(1)} \|^2 \leqslant 2nA_1B_1^2C = \mu_1 B_1 = 2B_2$$

Generally

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p)} \| \leqslant 2B_p \qquad (p = 0, 1, 2, \ldots) \tag{6}$$

Since on the basis of formula (2) of Sec. 13.4 the quantity $B_p \longrightarrow 0$ as $p \longrightarrow \infty$, then by passing to the limit in inequality (6) we get

$$\boldsymbol{x}^{**} = \lim_{p \to \infty} \boldsymbol{x}^{(p)} = \boldsymbol{x}^*$$

That is, the solution of system (1) in the domain $\| \boldsymbol{x} - \boldsymbol{x}^{(0)} \| \leqslant 2B_0$ is unique.

Note. If the domain $\overline{U}_{\mathscr{H}}(\boldsymbol{x}^{(0)})$ is such that

$$\frac{2}{\mu_0} B_0 \leqslant \mathscr{H}$$

then in the extended domain (1)

$$\| \boldsymbol{x} - \boldsymbol{x}^{(0)} \| \leqslant \frac{2}{\mu_0} B_0 \tag{7}$$

there are no solutions of system (1) other than $\boldsymbol{x}^*$.

Indeed, assuming that in the domain (7) there is a solution $\boldsymbol{x}^{**}$ of system (1) (Sec. 13.3) and repeating the reasoning of the theorem, we obtain an inequality of the form (3):

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p+1)} \| \leqslant \frac{1}{2} n A_p C \, \| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p)} \|^2$$

where $\boldsymbol{x}^{(p)}$ $(p = 0, 1, 2, \ldots)$ are successive approximations of the Newton process with initial approximation $\boldsymbol{x}^{(0)}$. From this, since

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(0)} \| \leqslant \frac{2}{\mu_0} B_0,$$

it follows that by using the numbers $\mu_{p+1} = \mu_p^2$ we get successively

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(1)} \| \leqslant \frac{1}{2} n A_0 C \, \frac{4}{\mu_0^2} B_0^2 =$$
$$= 2nA_0B_0C \cdot \frac{1}{\mu_0^2} B_0 = \frac{1}{\mu_0} B_0 = \frac{2}{\mu_0^2} B_1 = \frac{2}{\mu_1} B_1,$$
$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(2)} \| \leqslant \frac{1}{2} n A_1 C \cdot \frac{4}{\mu_1^2} B_1^2 = 2nA_1B_1C \cdot \frac{1}{2} \mu_1 B_1 \cdot \frac{2}{\mu_1^3} =$$
$$= \mu_1 \cdot B_2 \cdot \frac{2}{\mu_1^3} = \frac{2}{\mu_1^2} B_2 = \frac{2}{\mu_2} B_2$$

and so forth.

Generally

$$\| \boldsymbol{x}^{**} - \boldsymbol{x}^{(p)} \| \leqslant \frac{2}{\mu_p} B_p \qquad (p = 0, 1, 2, \ldots)$$

Since

$$B_p = \frac{1}{2}\mu_{p-1}B_{p-1}$$

and

$$\mu_p = \mu_{p-1}^2$$

then

$$\frac{B_p}{\mu_p} = \frac{1}{2} \cdot \frac{B_{p-1}}{\mu_{p-1}} = \left(\frac{1}{2}\right)^p \cdot \frac{B_0}{\mu_0} \tag{8}$$

The latter relation can also be obtained directly from the formulas (1) and (2) of Sec. 13.4.

Thus

$$\| x^{**} - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \frac{B_0}{\mu_0} \qquad (p=0, 1, 2, \ldots)$$

Hence

$$x^{**} = \lim_{p\to\infty} x^{(p)} = x^*$$

which is what we set out to prove.

*13.6 STABILITY OF CONVERGENCE OF THE NEWTON PROCESS UNDER VARIATIONS OF THE INITIAL APPROXIMATION

Theorem 4. *If Conditions* (1) *to* (4) *of Theorem* 1 (Sec. 13.3) *are fulfilled and*

$$\frac{2}{\mu_0} B_0 \leqslant \mathscr{H}$$

where $\mu_0 = 2nA_0B_0C < 1$, *then the Newton process converges to a unique solution* x^* *of the system* (1) *of* Sec. 13.3 *in the main domain* $\| x - x^{(0)} \| \leqslant 2B_0$ *for any choice of the initial approximation* $x^{(0)}$ *in the domain*

$$\| x'^{(0)} - x^{(0)} \| \leqslant \frac{1-\mu_0}{2\mu_0} B_0 \tag{1}$$

Proof. By analogy with the above-introduced notations

$$W_0 = W(x^{(0)}) \quad \text{and} \quad \Gamma_0 = W_0^{-1}$$

we introduce the designations

$$W_0' = W(x'^{(0)}) \quad \text{and} \quad \Gamma_0' = (W_0')^{-1}$$

We will show that conditions similar to the Conditions (1) to (4) of Theorem 1 will hold true at the point $x'^{(0)}$.

Using the notation and the method of proof of Theorem 1, we get

$$\| E - \Gamma_0 W_0' \| = \| \Gamma_0 (W_0 - W_0') \| \leqslant$$
$$\leqslant \| \Gamma_0 \| \| W_0 - W_0' \| \leqslant A_0 nC \| x'^{(0)} - x^{(0)} \|$$

whence, taking into account inequality (1), we obtain

$$\|E-\Gamma_0 W_0'\| \leqslant A_0 nC \frac{1-\mu_0}{2\mu_0} B_0 = \frac{1-\mu_0}{4} \leqslant \frac{1}{4}$$

Hence

$$\|(\Gamma_0 W_0')^{-1}\| = \|[E-(E-\Gamma_0 W_0')]^{-1}\| \leqslant$$

$$\leqslant \frac{1}{1-\|E-\Gamma_0 W_0'\|} \leqslant \frac{1}{1-\frac{1-\mu_0}{4}} = \frac{4}{3+\mu_0} \tag{2}$$

And so there exists

$$\Gamma_0' = (\Gamma_0 W_0')^{-1}\Gamma_0$$

and

$$\|\Gamma_0'\| \leqslant \|(\Gamma_0 W_0')^{-1}\| \, \|\Gamma_0\| \leqslant \frac{4A_0}{3+\mu_0} = A' \tag{3}$$

We then derive

$$\|\Gamma_0 \boldsymbol{f}(\boldsymbol{x}'^{(0)})\| \leqslant \|\Gamma_0\| \, \|\boldsymbol{f}(\boldsymbol{x}'^{(0)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) -$$

$$- W_0(\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)})\| + \|\Gamma_0 \boldsymbol{f}(\boldsymbol{x}^{(0)})\| + \|\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)}\| \leqslant$$

$$\leqslant \frac{1}{2} A_0 nC \|\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)}\|^2 + B_0 + \|\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)}\| \leqslant$$

$$\leqslant \frac{1}{4} \mu_0 B_0 \frac{1-2\mu_0+\mu_0^2}{4\mu_0^2} + B_0 + \frac{1-\mu_0}{2\mu_0} B_0 =$$

$$= \frac{1-2\mu_0+\mu_0^2+16\mu_0+8-8\mu_0}{16\mu^0} B_0 = \frac{(3+\mu_0)^2}{16\mu_0} B_0$$

From this, using inequality (2), we have

$$\|\Gamma_0' \boldsymbol{f}(\boldsymbol{x}'^{(0)})\| = \|(\Gamma_0 W_0')^{-1} \cdot \Gamma_0 \boldsymbol{f}(\boldsymbol{x}'^{(0)})\| \leqslant$$

$$\leqslant \|(\Gamma_0 W_0')^{-1}\| \cdot \|\Gamma_0 \boldsymbol{f}(\boldsymbol{x}'^{(0)})\| \leqslant$$

$$\leqslant \frac{4}{3+\mu_0} \cdot \frac{(3+\mu_0)^2}{16\mu_0} B_0 = \frac{3+\mu_0}{4\mu_0} B_0 = B' \tag{4}$$

On the basis of inequalities (3) and (4) we get

$$\mu' = 2nA'B'C = 2n \frac{4A_0}{3+\mu_0} \cdot \frac{3+\mu_0}{4\mu_0} B_0 C = 2nA_0 B_0 C \frac{1}{\mu_0} = 1$$

Besides,

$$2B' + \|\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)}\| \leqslant \frac{3+\mu_0}{2\mu_0} B_0 + \frac{1-\mu_0}{2\mu_0} B_0 = \frac{2B_0}{\mu_0} \leqslant \mathscr{H}$$

and hence all the more so

$$2B' \leqslant \frac{2B_0}{\mu_0} \leqslant \mathscr{H}$$

Thus, the conditions of Theorem 1 are completely fulfilled at the point $x'^{(0)}$, and

$$\overline{U}_{2B'}(x'^{(0)}) \subset \overline{U}_{\frac{2B_0}{\mu_0}}(x^{(0)}) \subset \overline{U}_{\mathscr{H}}(x^{(0)}) \tag{5}$$

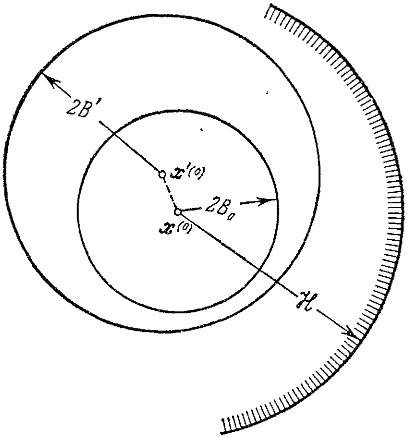


Fig. 58

For this reason, the Newton process

$$x'^{(p+1)} = x'^{(p)} - \Gamma'_p f(x'^{(p)})$$

where

$$\Gamma'_p = W^{-1}(x'^{(p)}) \qquad (p = 0, 1, 2, \ldots)$$

converges to a certain solution x'^* of system (1), Sec. 13.3, lying in the domain $\overline{U}_{2B'}(x'^{(0)})$. On the basis of formula (5)

$$x'^* \in \overline{U}_{\frac{2B_0}{\mu_0}}(x^{(0)})$$

But by virtue of the note pertaining to Theorem 3 of the preceding section there is a **unique solution** x^* of the basic system (1) in the domain $\overline{U}_{\frac{2B_0}{\mu_0}}(x^{(0)})$, and so

$$x'^* = x^*$$

and

$$x^* = \lim_{p \to \infty} x'^{(p)}$$

which completes the proof.

Note. If $2B_0 < \mathscr{H}$ and $\mu_0 < 1$, then for the initial approximation $x^{(0)}$ there is always a neighbourhood, any point of which can be taken for the initial approximation of a Newton process convergent to the desired solution $\boldsymbol{x}^*$.

Indeed, suppose

$$2B_0 < 2qB_0 = \mathcal{H}$$

where $q > 1$. Setting

$$\mu_0^* = \max\left(\mu_0, \frac{1}{q}\right)$$

we find that by Theorems 1 and 4 the appropriate Newton process, for any initial approximation $\boldsymbol{x}'^{(0)}$ satisfying the condition

$$\|\boldsymbol{x}'^{(0)} - \boldsymbol{x}^{(0)}\| \leqslant \frac{1-\mu_0^*}{2\mu_0} B_0$$

will converge to the solution $\boldsymbol{x}^*$ of system (1).

13.7 THE MODIFIED NEWTON METHOD

An essential inconvenience in forming the Newton process

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - W^{-1}(\boldsymbol{x}^{(p)})\boldsymbol{f}(\boldsymbol{x}^{(p)}) \qquad (p = 0, 1, 2, \ldots) \qquad (1)$$

is the necessity to compute the inverse matrix $W^{-1}(\boldsymbol{x}^{(p)})$ for each step. If the matrix $W^{-1}(\boldsymbol{x})$ is continuous in the neighbourhood of the desired solution $\boldsymbol{x}^*$ and the initial approximation $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{x}^*$, then we can approximately put

$$W^{-1}(\boldsymbol{x}^{(p)}) \approx W^{-1}(\boldsymbol{x}^{(0)})$$

We thus arrive at the *modified Newton process*

$$\xi^{(p+1)} = \xi^{(p)} - W^{-1}(\boldsymbol{x}^{(0)})\boldsymbol{f}(\xi^{(p)}) \qquad (2)$$

$(p = 0, 1, 2, \ldots)$, where $\xi^{(0)} = \boldsymbol{x}^{(0)}$. Note that the first approximations $\boldsymbol{x}^{(1)}$ and $\xi^{(1)}$ coincide for the processes (1) and (2):

$$\boldsymbol{x}^{(1)} = \xi^{(1)}$$

The convergence of the modified Newton process (2) has been investigated by L. V. Kantorovich [1].

Theorem. *If the Conditions* (1) *to* (4) *of Theorem* 1 (Sec. 13.3) *are fulfilled and*

$$\mu_0 = 2nA_0B_0C < 1$$

then the modified Newton process (2) *defined by the initial approximation* $\xi^{(0)} = \mathrm{x}^{(0)}$ *converges to the solution* x^* *of the system*

$$\boldsymbol{f}(\boldsymbol{x}) = 0$$

and

$$\|\boldsymbol{x}^* - \xi^{(p)}\| \leqslant \mu_0^p \|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\| \leqslant 2B_0\mu_0^p \qquad (p = 0, 1, 2, \ldots) \quad (3)$$

where the norm is understood to be the m-*norm.*

Proof. Consider the vector function

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} - \Gamma_0 \boldsymbol{f}(\boldsymbol{x}) = [F_i(\boldsymbol{x})]$$

where $\Gamma_0 = W^{-1}(\boldsymbol{x}^{(0)})$.

Obviously

$$\boldsymbol{F}(\xi^{(p)}) = \xi^{(p)} - \Gamma_0 \boldsymbol{f}(\xi^{(p)}) = \xi^{(p+1)} \qquad (p = 0, 1, 2, \ldots) \tag{4}$$

Moreover

$$\boldsymbol{F}'(\boldsymbol{x}) = E - \Gamma_0 \boldsymbol{f}'(\boldsymbol{x}) \tag{5}$$

whence, in particular,

$$\boldsymbol{F}'(\boldsymbol{x}^{(0)}) = E - \Gamma_0 \boldsymbol{f}'(\boldsymbol{x}^{(0)}) = E - E = 0 \tag{6}$$

We will prove by induction that all the approximations $\xi^{(p)}$ $(p = 1, 2, \ldots)$ lie in the $2B_0$-neighbourhood of the point $\boldsymbol{x}^{(0)}$, that is

$$\| \xi^{(p)} - \boldsymbol{x}^{(0)} \| < 2B_0 \tag{7}$$

Indeed, for $p = 1$, (7) is obvious since by Condition (2) of the theorem, we have

$$\| \xi^{(1)} - \boldsymbol{x}^{(0)} \| = \| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \leqslant B_0$$

Now suppose that for some p the inequality (7) is valid. Then, using Lemma 2 (Sec. 13.2), we have

$$\begin{aligned} \|\xi^{(p+1)} - \boldsymbol{x}^{(0)}\| &= \|\boldsymbol{F}(\xi^{(p)}) - \boldsymbol{x}^{(0)}\| = \|\xi^{(p)} - \Gamma_0 \boldsymbol{f}(\xi^{(p)}) - \boldsymbol{x}^{(0)}\| = \\ &= \|\Gamma_0 [\boldsymbol{f}(\xi^{(p)}) - W(\boldsymbol{x}^{(0)})(\xi^{(p)} - \boldsymbol{x}^{(0)})]\| \leqslant \|\Gamma_0 \boldsymbol{f}(\boldsymbol{x}^{(0)})\| + \\ &+ \|\Gamma_0 \{\boldsymbol{f}(\xi^{(p)}) - \boldsymbol{f}(\boldsymbol{x}^{(0)}) - W(\boldsymbol{x}^{(0)})(\xi^{(p)} - \boldsymbol{x}^{(0)})\}\| \leqslant \\ &\leqslant B_0 + \frac{1}{2} A_0 n C \|\xi^{(p)} - \boldsymbol{x}^{(0)}\|^2 \end{aligned}$$

Using (7), we find

$$\begin{aligned} \|\xi^{(p+1)} - \boldsymbol{x}^{(0)}\| &< B_0 + \frac{1}{2} n A_0 C \cdot 4B_0^2 = \\ &= B_0 + 2nA_0B_0C \cdot B_0 = (1 + \mu_0) B_0 < 2B_0 \end{aligned}$$

which proves our assertion.

Since the conditions of Theorem 1 of Sec. 13.3 are assumed to be fulfilled, the system $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ has a root $\boldsymbol{x}^*$ such that $\| \boldsymbol{x}^* - \boldsymbol{x}^{(0)} \| \leqslant 2B_0$.

Let us consider the difference $\boldsymbol{x}^* - \xi^{(p)}$, where $p \geqslant 1$. Taking into account that

$$\boldsymbol{F}(\boldsymbol{x}^*) \equiv \boldsymbol{x}^* - \Gamma_0 \boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{x}^*$$

and utilizing Lemma 1 of Sec. 13.2, we have

$$\| \boldsymbol{x}^* - \xi^{(p)} \| = \| \boldsymbol{F}(\boldsymbol{x}^*) - \boldsymbol{F}(\xi^{(p-1)}) \| \leqslant \| \boldsymbol{x}^* - \xi^{(p-1)} \| \cdot \| \boldsymbol{F}'(\theta) \| \tag{8}$$

where θ is a point in the interval $[\boldsymbol{x}^*, \xi^{(p-1)}]$.

Furthermore (see Sec. 13.2, Lemma 1, Corollary 2)

$$\|F'(\theta)\| = \|F'(\theta) - F'(x^{(0)})\| \leqslant n \|\theta - x^{(0)}\| \max \|F''(\eta)\| \tag{9}$$

where η is a point in the interval $[\theta, x^{(0)}]$. From formula (5) we have

$$F'(x) = \left[\delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j} \right]$$

where δ_{ij} is the Kronecker delta and $\Gamma_0 = [\gamma_{ij}]$. Therefore

$$\frac{\partial F_i}{\partial x_j} = \delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j}$$

and

$$\frac{\partial^2 F_i}{\partial x_j \partial x_k} = -\sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s}{\partial x_j \partial x_k}$$

Consequently

$$\|F''(\eta)\| + \max_{i,\,j} \sum_{s=1}^{n} \left| \frac{\partial^2 F_i(\eta)}{\partial x_j \partial x_k} \right| = \max_{i,\,j} \sum_{k=1}^{n} \left| \sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant$$

$$\leqslant \max_{i,\,j} \sum_{k=1}^{n} |\gamma_{is}| \sum_{k=1}^{n} \left| \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,\,j} \sum_{s=1}^{n} |\gamma_{is}| C = C \|\Gamma_0\| \leqslant A_0 C$$

and, hence, on the basis of (9),

$$\|F'(\theta)\| \leqslant n A_0 C \|\theta - x^{(0)}\|$$

Since the point θ plainly belongs to the $2B_0$-neighbourhood of the point $x^{(0)}$, it follows that

$$\|\theta - x^{(0)}\| \leqslant 2B_0$$

and, hence,

$$\|F'(\theta)\| \leqslant 2nA_0B_0C = \mu_0 \tag{10}$$

Taking into consideration inequality (10), we derive from inequality (8) that

$$\|x^* - \xi^{(p)}\| \leqslant \mu_0 \|x^* - \xi^{(p-1)}\|$$

whence

$$\|x^* - \xi^{(p)}\| \leqslant \mu_0^p \|x^* - \xi^{(0)}\| = \mu_0^p \|x^* - x^{(0)}\| \leqslant 2B\mu_0^p$$

From the last inequality there follows, for $\mu_0 < 1$,

$$\lim_{p \to \infty} \xi^{(p)} = x^*$$

The proof of the theorem is complete.

13.8 THE METHOD OF ITERATION

Given a system of nonlinear equations of a special kind

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{1}$$

where the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ are real, defined and continuous in some neighbourhood ω of an isolated solution $(x_1^*, x_2^*, \ldots, x_n^*)$ of this system.

Introducing the vectors

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad \varphi(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_n(\boldsymbol{x}))$$

we can write system (1) more compactly:

$$\boldsymbol{x} = \varphi(\boldsymbol{x}) \tag{2}$$

In finding the vector root $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ of equation (2), it is often convenient to use the *method of iteration*

$$\boldsymbol{x}^{(p+1)} = \varphi(\boldsymbol{x}^{(p)}) \qquad (p = 0, 1, 2, \ldots) \tag{3}$$

where the initial approximation $\boldsymbol{x}^{(0)} \approx \boldsymbol{x}^*$. The convergence of this process will be investigated below. Note that if the process of iteration (3) converges, then the limiting value

$$\xi = \lim_{p \to \infty} \boldsymbol{x}^{(p)} \tag{4}$$

is definitely a root of equation (2). Indeed, assuming that relation (4) is valid, and passing to the limit in (3) as $p \to \infty$, we will have, by virtue of the continuity of the function $\varphi(\boldsymbol{x})$,

$$\lim_{p \to \infty} \boldsymbol{x}^{(p+1)} = \varphi\left(\lim_{p \to \infty} \boldsymbol{x}^{(p)}\right)$$

that is,

$$\xi = \varphi(\xi)$$

Thus, ξ is a root of the vector equation (2).

If, besides, all the approximations $\boldsymbol{x}^{(p)}$ $(p = 0, 1, 2, \ldots)$ belong to the domain ω and $\boldsymbol{x}^*$ is a **unique** root of system (2) in ω, then, clearly,

$$\xi = \boldsymbol{x}^*$$

The method of iteration can also be applied to the general system

$$\boldsymbol{f}(\boldsymbol{x}) = 0 \tag{5}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is a vector function defined and continuous in the neighbourhood ω of an isolated vector root $\boldsymbol{x}^*$. For example, rewrite this system as

$$x = x + \Lambda f(x)$$

where Λ is a nonsingular matrix. Introducing the notation

$$x + \Lambda f(x) = \varphi(x) \tag{6}$$

we will have

$$x = \varphi(x) \tag{7}$$

The ordinary method of iteration (3) is readily applicable to this equation.

If the function $f(x)$ has a continuous derivative $f'(x)$ in ω, then from formula (6) follows

$$\varphi'(x) = E + \Lambda f'(x)$$

In the sections which follow, proof will be given that the process of iteration rapidly converges for equation (7) if $\varphi'(x)$ is small in norm. Taking this circumstance into account, we choose matrix Λ so that

$$\varphi'(x^{(0)}) = E + \Lambda f'(x^{(0)}) = 0$$

whence, if the matrix $f'(x^{(0)})$ is nonsingular, we will have

$$\Lambda = -[f'(x^{(0)})]^{-1}$$

Note that essentially this is the modified Newton process applied to equation (5) (see Sec. 13.7).

If $\det f'(x^{(0)}) = 0$, then one should choose a different initial approximation $x^{(0)}$.

There are also other ways of replacing system (5) by the equivalent system (7).

Example. Use the method of iteration to give an approximate solution of the system

$$\left.\begin{aligned} x_1^2 + x_2^2 &= 1, \\ x_1^3 - x_2 &= 0 \end{aligned}\right\} \tag{8}$$

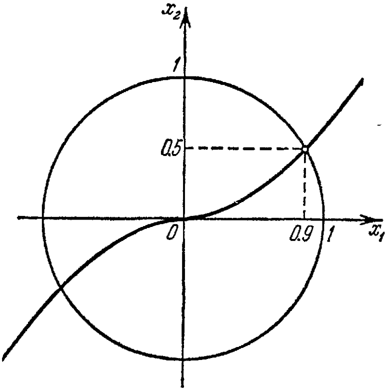


Fig. 59

Solution. From a graphical construction (Fig. 59), it can be seen that the system (8) has two solutions that differ only in sign. We limit ourselves to the positive solution. From the figure we see that it is possible to take

$$x^{(0)} = \begin{bmatrix} 0.9 \\ 0.5 \end{bmatrix}$$

for the initial approximation of the positive solution of system (8).

Setting

$$\boldsymbol{f}(\boldsymbol{x})=\begin{bmatrix} x_1^2+x_2^2-1 \\ x_1^3-x_2 \end{bmatrix}$$

we have

$$\boldsymbol{f}'(\boldsymbol{x})=\begin{bmatrix} 2x_1 & 2x_2 \\ 3x_1^2 & -1 \end{bmatrix}$$

whence

$$\boldsymbol{f}'(\boldsymbol{x}^{(0)})=\begin{bmatrix} 1.8 & 1 \\ 2.43 & -1 \end{bmatrix}$$

and

$$\det \boldsymbol{f}'(\boldsymbol{x}^{(0)})=-1.8-2.43=-4.23$$

Since the matrix $\boldsymbol{f}'(\boldsymbol{x}^{(0)})$ is nonsingular, there is an inverse matrix

$$[\boldsymbol{f}'(\boldsymbol{x}^{(0)})]^{-1}=-\frac{1}{4.23}\begin{bmatrix} -1 & -1 \\ -2.43 & 1.8 \end{bmatrix}$$

Thus

$$\Lambda=-[\boldsymbol{f}'(\boldsymbol{x}^{(0)})]^{-1}=\frac{1}{4.23}\begin{bmatrix} -1 & -1 \\ -2.43 & 1.8 \end{bmatrix}$$

Put

$$\varphi(\boldsymbol{x})=\boldsymbol{x}+\Lambda\boldsymbol{f}(\boldsymbol{x})=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}-\frac{1}{4.23}\begin{bmatrix} 1 & 1 \\ 2.43 & -1.8 \end{bmatrix}\begin{bmatrix} x_1^2+x_2^2-1 \\ x_1^3-x_2 \end{bmatrix}$$

Then the system (8) will be equivalent to the standard matrix equation

$$\boldsymbol{x}=\varphi(\boldsymbol{x}) \tag{9}$$

Using formula (4), we find the following successive approximations for the solution of system (9):

$$\boldsymbol{x}^{(1)}=\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix}-\frac{1}{4.23}\begin{bmatrix} 1 & 1 \\ 2.43 & 1.8 \end{bmatrix}\begin{bmatrix} x_1^{(0)^2}+x_2^{(0)^2}-1 \\ x_1^{(0)^3}-x_2^{(0)} \end{bmatrix}=$$

$$=\begin{bmatrix} 0.9 \\ 0.5 \end{bmatrix}-\frac{1}{4.23}\begin{bmatrix} 1 & 1 \\ 2.43 & -1.8 \end{bmatrix}\begin{bmatrix} 0.060 \\ 0.229 \end{bmatrix}=$$

$$=\begin{bmatrix} 0.9 \\ 0.5 \end{bmatrix}-\begin{bmatrix} 0.0683 \\ -0.0630 \end{bmatrix}=\begin{bmatrix} 0.8317 \\ 0.5630 \end{bmatrix},$$

$$\boldsymbol{x}^{(2)}=\begin{bmatrix} 0.8317 \\ 0.5630 \end{bmatrix}-\frac{1}{4.23}\begin{bmatrix} 1 & 1 \\ 2.43 & -1.8 \end{bmatrix}\begin{bmatrix} 0.8317^2+0.5630^2-1 \\ 0.8317^3-0.5630 \end{bmatrix}=$$

$$= \begin{bmatrix} 0.8317 \\ 0.5630 \end{bmatrix} - \begin{bmatrix} -0.0049 \\ 0.0003 \end{bmatrix} = \begin{bmatrix} 0.8268 \\ 0.5633 \end{bmatrix},$$

$$\boldsymbol{x}^{(3)} = \begin{bmatrix} 0.8268 \\ 0.5633 \end{bmatrix} - \begin{bmatrix} 0.0007 \\ 0.0002 \end{bmatrix} = \begin{bmatrix} 0.8261 \\ 0.5631 \end{bmatrix},$$

$$\boldsymbol{x}^{(4)} = \begin{bmatrix} 0.8261 \\ 0.5631 \end{bmatrix} - \begin{bmatrix} 0.0000 \\ -0.0005 \end{bmatrix} = \begin{bmatrix} 0.8261 \\ 0.5636 \end{bmatrix}$$

and so on.

Stopping with the fourth approximation, we have the roots

$$x_1 = 0.8261, \quad x_2 = 0.5636$$

and

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 0.0000 \\ 0.0002 \end{bmatrix}$$

to within an accuracy of 10^{-4}.

*13.9 THE NOTION OF A CONTRACTION MAPPING

Let there be given a nonlinear system

$$\left.\begin{array}{l} y_1 = \varphi_1 (x_1, x_2, \ldots, x_n), \\ y_2 = \varphi_2 (x_1, x_2, \ldots, x_n), \\ \cdots\cdots\cdots\cdots \\ y_n = \varphi_n (x_1, x_2, \ldots, x_n) \end{array}\right\} \tag{1}$$

where the functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ are defined and continuous in a known domain G of a real n-dimensional space E_n (Sec. 10.1), their values $(y_1, y_2, \ldots, y_n)$ for $(x_1, x_2, \ldots, x_n) \in G$ filling some domain $G' \subset E_n$. System (1) establishes a *mapping* of domain G onto G'. Introducing the vectors

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 \\ \cdot \\ \cdot \\ \cdot \\ \varphi_n \end{bmatrix}$$

we can write (1) compactly as

$$\boldsymbol{y} = \boldsymbol{\varphi}(\boldsymbol{x}) \tag{1'}$$

We introduce in space E_n the canonical norm $\|\boldsymbol{x}\|$ which satisfies the ordinary conditions. For example, one can put

$$\|\boldsymbol{x}\|_m = \max_i |x_i|$$

or

$$\|\boldsymbol{x}\|_l = \sum_i |x_i|$$

or

$$\|\boldsymbol{x}\|_k = \sqrt{\sum_i x_i^2}$$

and so on.

The mapping (1) or (1′) is termed a *contraction* mapping in the domain G if there exists a proper fraction q such that for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in G$ their images $\boldsymbol{y}_1 = \varphi(\boldsymbol{x}_1)$ and $\boldsymbol{y}_2 = \varphi(\boldsymbol{x}_2)$ satisfy the condition

$$\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \leqslant q \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \tag{2}$$

That is,

$$\|\varphi(\boldsymbol{x}_1) - \varphi(\boldsymbol{x}_2)\| \leqslant q \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \qquad (0 \leqslant q < 1) \tag{2′}$$

Let us consider the nonlinear vector equation

$$\boldsymbol{x} = \varphi(\boldsymbol{x}) \tag{3}$$

which is equivalent to the special kind of nonlinear system of equations

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{3′}$$

The solution $\boldsymbol{x}^*$ of this equation, if it exists, is a *fixed point* of the transformation (1). To find $\boldsymbol{x}^*$, construct the *iteration process*

$$\boldsymbol{x}^{(p)} = \varphi(\boldsymbol{x}^{(p-1)}) \qquad (p = 1, 2, \ldots) \tag{4}$$

where $\boldsymbol{x}^{(0)} \in G$.

Theorem 1. *Let domain* G *be closed and let the mapping* (1) *be a contraction mapping in* G; *that is, condition* (2) *is fulfilled. Then, if for the iteration process* (4) *all successive approximations* $\mathrm{x}^{(p)} \in \mathrm{G}$ $(\mathrm{p} = 0, 1, 2, \ldots)$, *it follows that* (1) *irrespective of the choice of the initial approximation* $\mathrm{x}^{(0)}$ *the process* (4) *converges, i. e., there exists the limit*

$$\boldsymbol{x}^* = \lim_{p \to \infty} \boldsymbol{x}^{(p)} \tag{5}$$

(2) *the limiting vector* x^* *is the sole solution of equation* (3) *in the domain* G; (3) *the estimate*

$$\|\boldsymbol{x}^* - \boldsymbol{x}^{(p)}\| \leqslant \frac{q^p}{1-q} \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\| \tag{6}$$

holds true.

Proof. (1) To prove the convergence of the sequence of approximations $x^{(p)}$ $(p=0, 1, 2, \ldots)$, we apply the *Cauchy test* (see Sec. 7.9) to get

$$\| x^{(p+k)} - x^{(p)} \| = \| (x^{(p+1)} - x^{(p)}) + (x^{(p+2)} - x^{(p+1)}) + \ldots \\ \ldots + (x^{(p+k)} - x^{(p+k-1)}) \| \leqslant \| x^{(p+1} - x^{(p)} \| + \\ + \| x^{(p+2)} - x^{(p+1)} \| + \ldots + \| x^{(p+k)} - x^{(p+k-1)} \| \tag{7}$$

Utilizing relation (4) and the "contraction condition" (2′), we successively obtain

$$\| x^{(s+1)} - x^{(s)} \| = \| \varphi(x^{(s)}) - \varphi(x^{(s-1)}) \| \leqslant \\ \leqslant q \| x^{(s)} - x^{(s-1)} \| \leqslant q^2 \| x^{(s-1)} - x^{(s-2)} \| \leqslant \\ \leqslant q^s \| x^{(1)} - x^{(0)} \| \tag{8}$$

where $s \geqslant 0$. Therefore, strengthening the right member of inequality (7), we get

$$\| x^{(p+k)} - x^{p} \| \leqslant q^p \| x^{(1)} - x^{(0)} \| + q^{p+1} \| x^{(1)} - x^{(0)} \| + \ldots \\ + q^{p+k-1} \| x^{(1)} - x^{(0)} \|$$

or, using the formula for the sum of terms of a geometric progression, we find

$$\| x^{(p+k)} - x^{(p)} \| \leqslant \frac{q^p - q^{p+k}}{1-q} \| x^{(1)} - x^{(0)} \| \leqslant \frac{q^p}{1-q} \| x^{(1)} - x^{(0)} \| \tag{9}$$

Since $0 \leqslant q < 1$ and, hence, $q^p \to 0$ as $p \to \infty$, from formula (9) it follows that for any $\varepsilon > 0$ there is an $N = N(\varepsilon)$ such that for $p > N(\varepsilon)$ and $k > 0$ the inequality

$$\| x^{(p+k)} - x^{(p)} \| < \varepsilon$$

holds true, that is, for the sequence $x^{(p)}$ $(p=0, 1, 2, \ldots)$ Cauchy's test is valid. Therefore,

$$x^* = \lim_{p \to \infty} x^{(p)}$$

and $x^* \in G$ since the domain G is closed.

(2) The vector x^* is a solution of equation (3) since, by passing to the limit in (4) as $p \to \infty$ and taking into account the continuity in G of the vector function $\varphi(x)$, we have

$$\lim_{p \to \infty} x^{(p)} = \varphi(\lim_{p \to \infty} x^{(p-1)})$$

that is

$$x^* = \varphi(x^*) \tag{10}$$

This solution is unique in G. Indeed, let $x^{*\prime}$ be another solution of (3):

$$x^{*\prime} = \varphi(x^{*\prime}) \tag{11}$$

Subtracting (11) from (10), we get

$$x^* - x^{*\prime} = \varphi(x^*) - \varphi(x^{*\prime})$$

whence

$$\| x^* - x^{*\prime} \| = \| \varphi(x^*) - \varphi(x^{*\prime}) \| \leqslant q \| x^* - x^{*\prime} \|$$

or

$$(1-q) \| x^* - x^{*\prime} \| \leqslant 0 \tag{12}$$

Since $1-q > 0$, then inequality (12) can only hold if $\| x^* - x^{*\prime} \| = 0$, that is, when $x^* = x^{*\prime}$. Thus, there cannot be any other solution of equation (3) in the domain G.

(3) Passing to the limit in inequality (9) as $k \to \infty$ we get the estimate (6).

This completes the proof of Theorem 1.

Note 1. If G coincides with the whole space E_n, then the condition $x^{(p)} \in G$ $(p = 0, 1, 2, \ldots)$ is obviously superfluous.

Note 2. Using the inequalities

$$\| x^{(p+1)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \|,$$
$$\| x^{(p+2)} - x^{(p+1)} \| \leqslant q^2 \| x^{(p)} - x^{(p-1)} \|,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$

we obtain from formula (7)

$$\| x^{(p+k)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \| + q^2 \| x^{(p)} - x^{(p-1)} \| +$$
$$+ \ldots + q^k \| x^{(p)} - x^{(p-1)} \| \leqslant \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|$$

whence, as $k \to \infty$, we get

$$\| x^* - x^{(p)} \| \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|. \tag{13}$$

In particular, if $0 \leqslant q \leqslant \frac{1}{2}$, then formula (13) implies that for

$$\| x^{(p)} - x^{(p-1)} \| \leqslant \varepsilon$$

the inequality

$$\| x^* - x^{(p)} \| \leqslant \varepsilon$$

holds true.

Under the hypothesis of Theorem 1 it is required that **all** the approximations $x^{(p)}$ belong to the fixed domain G. This is sometimes hard to verify. For this reason, we will give a slightly modified statement of Theorem 1.

Theorem 2. *Let* (1) *be a contraction mapping in a closed domain* G *and let* g *be a bounded domain lying in* G *together with its* ρ*-neighbourhood* (in the sense of the norm that has been introduced), *where*

$$\rho \geqslant \frac{Dq}{1-q} \tag{14}$$

Furthermore, let D *be the diameter of* g *and let* q *be a corresponding coefficient in the inequality* (2). *Then, if*

$$\boldsymbol{x}^{(0)} \in g \quad \text{and} \quad \boldsymbol{x}^{(1)} = \varphi(\boldsymbol{x}^{(0)}) \in g$$

the iteration process (4) *converges and the conclusions of Theorem* 1 *hold true.*

Proof. It obviously suffices to prove that all the successive approximations lie in the domain G:

$$\boldsymbol{x}^{(p)} \in G \qquad (p = 2, 3, \ldots) \tag{15}$$

We carry out the proof by induction.

First of all, by hypothesis, $\boldsymbol{x}^{(0)} \in G$ and $\boldsymbol{x}^{(1)} \in G$. Now let $\boldsymbol{x}^{(s)} \in G\, (s = 0, 1, 2, \ldots, p)$. Since

$$\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \leqslant D$$

we have, on the basis of inequality (8),

$$\| \boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)} \| \leqslant q^s \| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \leqslant Dq^s \,(s = 0, 1, 2, \ldots, p)$$

Therefore

$$\| \boldsymbol{x}^{(p+1} - \boldsymbol{x}^{(1)} \| \leqslant \| \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \| + \ldots + \| x^{(p+1)} - x^{(p)} \| \leqslant$$
$$\leqslant Dq + \ldots + Dq^p \leqslant \frac{Dq}{1-q} \leqslant \rho$$

Consequently, $\boldsymbol{x}^{(p+1)} \in G$.

Thus, Property (15) is established and we are in the conditions of Theorem 1. The proof of Theorem 2 is complete.

Note. The conditions of Theorem 2 are equivalent to the following: let $\boldsymbol{x}^{(0)} \in G$ and $\boldsymbol{x}^{(1)} \in G$, where $\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \| \leqslant D$. If the distance of point $\boldsymbol{x}^{(1)}$ from the boundary Γ of G does not exceed $\frac{Dq}{1-q}$, then the iteration process (4) converges.

*13.10 FIRST SUFFICIENT CONDITION FOR THE CONVERGENCE OF THE PROCESS OF ITERATION

Let us consider a real reduced system

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{1}$$

or, in matrix form,

$$\boldsymbol{x} = \varphi(\boldsymbol{x}) \tag{1'}$$

where

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \varphi(\boldsymbol{x})=\begin{bmatrix} \varphi_1(\boldsymbol{x}) \\ \cdot \\ \cdot \\ \cdot \\ \varphi_n(\boldsymbol{x}) \end{bmatrix}$$

It is assumed that the vector function $\varphi(\boldsymbol{x})$ is defined and continuous together with its derivative $\varphi'(\boldsymbol{x})=\left[\frac{\partial\varphi_i}{\partial x_j}\right]$ in a convex, bounded and closed domain $G\subset E_n$.

In this section we make use of two norms:

$$\|\boldsymbol{x}\|_m=\max_i |x_i|$$

and

$$\|\boldsymbol{x}\|_l=\sum_{i=1}^{n}|x_i|$$

With respect to domain G we introduce the norms

$$\|\varphi'(\boldsymbol{x})\|_{\mathrm{I}}=\max_{x\in G}\|\varphi'(\boldsymbol{x})\|_m \tag{2}$$

and

$$\|\varphi'(\boldsymbol{x})\|_{\mathrm{II}}=\max_{x\in G}\|\varphi'(\boldsymbol{x})\|_l \tag{3}$$

where

$$\|\varphi'(\boldsymbol{x})\|_m=\max_i\sum_{j=1}^{n}\left|\frac{\partial\varphi_i(x)}{\partial x_j}\right| \tag{2'}$$

and

$$\|\varphi'(\boldsymbol{x})\|_l=\max_j\sum_{i=1}^{n}\left|\frac{\partial\varphi_i(x)}{\partial x_j}\right| \tag{3'}$$

Theorem. *Let the functions* $\varphi(\mathrm{x})$ *and* $\varphi'(\mathrm{x})$ *be continuous in the domain* G, *and, in* G, *let the inequality*

$$\|\varphi'(\boldsymbol{x})\|_{\mathrm{I}}\leqslant q<1, \tag{4}$$

where q *is a constant, hold true.*

If the successive approximations

$$\boldsymbol{x}^{(p+1)}=\varphi(\boldsymbol{x}^{(p)}) \tag{5}$$

$(\mathrm{p}=0, 1, 2\ldots)$ *lie in* G, *then the iteration process* (5) *converges and the limiting vector*

$$\boldsymbol{x}^{*}=\lim_{p\to\infty}\boldsymbol{x}^{(p)}$$

is the sole solution of system (1) *in domain* G.

Proof. By virtue of Theorem 1 of the preceding section, it suffices to demonstrate that the mapping

$$\boldsymbol{y}=\varphi(\boldsymbol{x}) \tag{6}$$

is, given Condition (4), a contraction mapping in G in the sense of the m-norm.

Suppose $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\in G$ and $\boldsymbol{y}_i=\varphi(\boldsymbol{x}_i)$ $(i=1,\ 2)$.
By the Corollary 1 of Lemma 1 (Sec. 13.2), we have

$$\|\boldsymbol{y}_1-\boldsymbol{y}_2\|_m=\|\varphi(\boldsymbol{x}_1)-\varphi(\boldsymbol{x}_2)\|_m\leqslant$$
$$\leqslant\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_m\|\varphi'(\xi)\|_m\leqslant\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_m\|\varphi'(\boldsymbol{x})\|_{\mathrm{I}}$$

whence

$$\|\boldsymbol{y}_1-\boldsymbol{y}_2\|_m\leqslant q\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_m$$

where $0\leqslant q<1$, which completes the proof.

Corollary. The iteration process (5) converges if

$$\sum_{j=1}^{n}\left|\frac{\partial\varphi_i(\boldsymbol{x})}{\partial x_j}\right|\leqslant q_i<1 \qquad (i=1,\ 2,\ \ldots,\ n) \tag{7}$$

when $\boldsymbol{x}\in G$.

Obviously, from the system of inequalities (7) follows Condition (4) of the theorem.

Note. On the basis of Theorem 1, Sec. 13.9, we obtain the following estimate for the approximation $\boldsymbol{x}^{(p)}$:

$$\|\boldsymbol{x}^*-\boldsymbol{x}^{(p)}\|_m\leqslant\frac{q^{(p)}}{1-q}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|_m \qquad (p=0,\ 1,\ 2.\ \ldots,)$$

where $\boldsymbol{x}^{(1)}-\varphi(\boldsymbol{x}^{(0)})$.

*13.11 SECOND SUFFICIENT CONDITION FOR THE CONVERGENCE OF THE PROCESS OF ITERATION

Before taking up the proof of the convergence theorem that uses the norm $\|\varphi'(\boldsymbol{x})\|_{\mathrm{II}}$ let us derive an estimate for the difference of the values of the vector function. This estimate is similar to the mean-value theorem and is also of interest in itself.

Lemma. *If the vector function*

$$f(\boldsymbol{x})=\begin{bmatrix}f_1(\boldsymbol{x})\\ \cdot\\ \cdot\\ \cdot\\ f_n(\boldsymbol{x})\end{bmatrix}$$

is continuous together with its derivative $\mathbf{f}'(\mathbf{x})$ *in a convex domain containing the points* $\mathbf{x}$ *and* $\mathbf{x}+\Delta\mathbf{x}$, *then*

$$\|f(x+\Delta x)-f(x)\|_l \leqslant \|\Delta x\|_l \cdot \|f'(\xi)\|_l \tag{1}$$

where $\xi=\mathbf{x}+\theta\Delta\mathbf{x}$ *and* $0<\theta<1$.

Proof. Consider the auxiliary function

$$\Phi(t)=\sum_{i=1}^{n}\varepsilon_i[f_i(x+t\Delta x)-f_i(x)]$$

where $0\leqslant t\leqslant 1$ is a scalar argument and ε_i is a sequence of numbers assuming the values -1, 0, 1. Clearly, $\Phi(0)=0$. Applying the mean-value theorem, we get

$$\sum_{i=1}^{n}\varepsilon_i[f_i(x+\Delta x)-f_i(x)]=\Phi(1)-\Phi(0)=\Phi'(\theta)=$$

$$=\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{n}\frac{\partial f_i(\xi)}{\partial x_j}\Delta x_j$$

where $\xi=x+\theta\Delta x$ and $0<\theta<1$.

From this we have, noting that $|\varepsilon_i|\leqslant 1$,

$$\sum_{i=1}^{n}\varepsilon_i[f_i(x+\Delta x)-f_i(x)]\leqslant$$

$$\leqslant\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{\partial f_i(\xi)}{\partial x_j}\right|\cdot|\Delta x_j|=\sum_{j=1}^{n}|\Delta x_j|\sum_{i=1}^{n}\left|\frac{\partial f_i(\xi)}{\partial x_j}\right| \tag{2}$$

Since

$$\sum_{i=1}^{n}\left|\frac{\partial f_i(\xi)}{\partial x_j}\right|\leqslant\max_j\sum_{i=1}^{n}\left|\frac{\partial f_i(\xi)}{\partial x_j}\right|=\|f'(\xi)\|_l$$

it follows that, by strengthening inequality (2), we get

$$\sum_{i=1}^{n}\varepsilon_i[f_i(x+\Delta x)-f_i(x)]\leqslant\sum_{j=1}^{n}|\Delta x_j|\,\|f'(\xi)\|_l=$$

$$=\|f'(\xi)\|_l\cdot\sum_{j=1}^{n}|\Delta x_j|=\|f'(\xi)\|_l\cdot\|\Delta x\|_l$$

Assuming, in the last inequality, that

$$\varepsilon_i=\operatorname{sgn}[f_i(x+\Delta x)-f_i(x)]\qquad(i=1,\ 2,\ \ldots,\ n)$$

we finally get

$$\sum_{i=1}^{n}|f_i(x+\Delta x)-f_i(x)|\leqslant\|f'(\xi)\|_l\,\|\Delta x\|_l$$

Thus,

$$\|f(x+\Delta x)-f(x)\|_l \leqslant \|\Delta x\|_l \|f'(\xi)\|_l \tag{2'}$$

which completes the proof.[1]

Theorem. *Let the vector function* $\varphi(x)$ *be continuous together with its derivative* $\varphi'(x)$ *in a bounded convex closed domain* G *and*

$$\|\varphi(x)\|_{\text{II}} \leqslant q < 1 \tag{3}$$

where q *is a constant. If* $x^{(0)} \in G$ *and all successive approximations*

$$x^{(p+1)} = \varphi(x^{(p)}) \qquad (p=0,\ 1,\ 2,\ \ldots) \tag{4}$$

also lie in G, *then the iteration process* (4) *converges to a unique solution of the equation*

$$x = \varphi(x) \tag{5}$$

in G

Proof. We will prove that $y=\varphi(x)$ is a contraction mapping in G in the sense of the l-norm.

Let $x_1,\ x_2 \in G$ and $y_i=\varphi(x_i)$ $(i=1,\ 2)$. Using the lemma, we have

$$\|y_1-y_2\|_l = \|\varphi(x_1)-\varphi(x_2)\|_l \leqslant \|x_1-x_2\|_l \cdot \|\varphi'(\xi)\|_l \tag{6}$$

where $\xi \in G$.

Since

$$\|\varphi'(\xi)\|_l \leqslant \max_{x \in G} \|\varphi'(x)\|_l = \|\varphi'(x)\|_{\text{II}} \leqslant q$$

it follows from inequality (6) that

$$\|y_1-y_2\|_l \leqslant q\,\|x_1-x_2\|_l$$

where $0 \leqslant q < 1$.

By the Theorem of Sec. 13.10, the proof is complete.

Corollary. The process of iteration (4) converges to the sole solution of equation (5) if the inequalities

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_j(x)}{\partial x_i} \right| \leqslant q_x < 1 \tag{7}$$

$(i=1,\ 2,\ldots,\ n)$ hold true for $x \in G$.

---

[1] If we apply the mean-value theorem directly to each component of the vector $f(x+\Delta x)-f(x)$, then the resulting estimate is dependent on the values of the derivative $\frac{\partial f_i(\xi_i)}{\partial x_j}$ at various points ξ_i $(i=1,\ 2,\ \ldots,\ n)$ of the interval $(x,\ x+\Delta x)$. Inequality (2') shows that we can confine ourselves to the values of the derivatives $\frac{\partial f_i(\xi)}{\partial x_j}$ at one and the same point $\xi \in (x,\ x+\Delta x)$.

Note. On the basis of the theorem of Sec. 13.10, we have, for the approximation $\boldsymbol{x}^{(p)}$, the following estimate

$$\| \boldsymbol{x}^* - \boldsymbol{x}^{(p)} \|_l \leqslant \frac{q^p}{1-q} \| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \|_l$$

where $\boldsymbol{x}^{(1)} = \varphi(\boldsymbol{x}^{(0)})$.

*13.12 THE METHOD OF STEEPEST DESCENT (GRADIENT METHOD)

Suppose we have a system of equations

$$\left.\begin{array}{l} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{array}\right\} \tag{1}$$

or, in matrix form,

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \tag{2}$$

$$\text{where } \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \\ f_n \end{bmatrix}.$$

Let us suppose that the functions f_i are real and continuously differentiable in their common domain of definition. Consider the function

$$U(\boldsymbol{x}) = \sum_{i=1}^{n} [f_i(\boldsymbol{x})]^2 = (\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})) \tag{3}$$

It is obvious that each solution of system (1) reduces the function $U(\boldsymbol{x})$ to zero; contrariwise, the numbers $x_1, x_2, \ldots, x_n$ for which the function $U(\boldsymbol{x})$ is zero are roots of the system (1).

We will assume that the system (1) has only an isolated solution which is the point of the strict minimum of $U(\boldsymbol{x})$. The problem thus reduces to finding the minimum of the function $U(\boldsymbol{x})$ in the n-dimensional space $E_n = \{x_1, x_2, \ldots, x_n\}$.

Let $\boldsymbol{x}$ be a vector root of (1) and let $\boldsymbol{x}^{(0)}$ be its zeroth approximation. Through the point $\boldsymbol{x}^{(0)}$ draw the level surface of the function $U(\boldsymbol{x})$. If point $\boldsymbol{x}^{(0)}$ is sufficiently close to the root $\boldsymbol{x}$, then, under our assumptions, the level surface

$$U(\boldsymbol{x}) = U(\boldsymbol{x}^{(0)})$$

will resemble an ellipsoid.

Start out from point $\boldsymbol{x}^{(0)}$ and move along the normal to the surface $U(\boldsymbol{x}) = U(\boldsymbol{x}^{(0)})$ until this normal touches some other level surface (Fig. 60)

$$U(\boldsymbol{x}) = U(\boldsymbol{x}^{(1)})$$

at some point $\boldsymbol{x}^{(1)}$.

Then, starting from point $\boldsymbol{x}^{(1)}$, move again along the normal to the level surface $U(\boldsymbol{x}) = U(\boldsymbol{x}^{(1)})$ until this normal touches at some point $\boldsymbol{x}^{(2)}$, the new level surface $U(\boldsymbol{x}) = U(\boldsymbol{x}^{(2)})$, etc.

Since $U(\boldsymbol{x}^{(0)}) > U(\boldsymbol{x}^{(1)}) > U(\boldsymbol{x})^{(2)} > \ldots$, by following this route we rapidly approach the point with the smallest value of U (the bottom of the "well") which corresponds to the required root $\boldsymbol{x}$ of the system (1). Denote the gradient[1] of the function $U(\boldsymbol{x})$ by

$$\nabla U(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial U}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial U}{\partial x_n} \end{bmatrix}$$

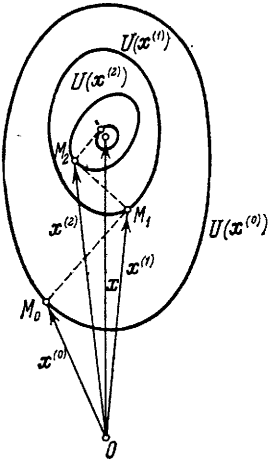


Fig. 60

From the vector triangles OM_0M_1, OM_1M_2, ... we conclude that

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \lambda_p \nabla U(\boldsymbol{x}^{(p)}) \qquad (p = 0, 1, 2, \ldots)$$

It remains to determine the factors λ_p. To do this, consider the scalar function

$$\Phi(\lambda) = U[\boldsymbol{x}^{(p)} - \lambda \nabla U(\boldsymbol{x}^{(p)})]$$

The function $\Phi(\lambda)$ gives the variation of the level of the function U along the appropriate normal to the level surface at the point $\boldsymbol{x}^{(p)}$. The factor $\lambda = \lambda_p$ must be chosen so that $\Phi(\lambda)$ has

---

[1] The gradient of the function $U(\boldsymbol{x})$ (it is denoted by grad U or ∇U, where the symbol ∇ is called *nabla* or *del*) is the vector applied to point $\boldsymbol{x}$ and having the direction of the normal $\boldsymbol{n}$ to the level surface of the function at the given point towards increasing U and having length equal to $\frac{\partial U}{\partial n}$.

The following formula is valid:

$$\nabla U = \frac{\partial U}{\partial x_1} \boldsymbol{e}_1 + \frac{\partial U}{\partial x_2} \boldsymbol{e}_2 + \ldots + \frac{\partial U}{\partial x_n} \boldsymbol{e}_n$$

where $\boldsymbol{e}_i$ $(i = 1, 2, \ldots, n)$ are the unit vectors of space E_n.

a minimum. Taking the derivative with respect to λ and equating it to zero, we obtain the equation

$$\Phi'(\lambda)=\frac{\partial}{\partial\lambda}U\left[\boldsymbol{x}^{(p)}-\lambda\nabla U(\boldsymbol{x}^{(p)})\right]=0 \tag{4}$$

The least positive root of equation (4) yields the value of λ_p. Equation (4) must, generally speaking, be solved numerically, and so we give a method for approximating the numbers λ_p. We will assume that λ is a small quantity, the square and higher powers of which may be neglected. We have

$$\Phi(\lambda)=\sum_{i=1}^{n}\{f_i[\boldsymbol{x}^{(p)}-\lambda\nabla U(\boldsymbol{x}^{(p)})]\}^2$$

Expanding the functions f_i in powers of λ to within linear terms, we obtain

$$\Phi(\lambda)=\sum_{i=1}^{n}\left[f_i(\boldsymbol{x}^{(p)})-\lambda\frac{\partial f_i(\boldsymbol{x}^{(p)})}{\partial\boldsymbol{x}}\nabla U(\boldsymbol{x}^{(p)})\right]^2$$

where

$$\frac{\partial f_i}{\partial x}=\left[\frac{\partial f_i}{\partial x_1},\ \frac{\partial f_i}{\partial x_2},\ \ldots,\ \frac{\partial f_i}{\partial x_n}\right]$$

Whence

$$\Phi'(\lambda)=-2\sum_{i=1}^{n}\left[f_i(\boldsymbol{x}^{(p)})-\lambda\frac{\partial f_i(\boldsymbol{x}^{(p)})}{\partial\boldsymbol{x}}\nabla U(\boldsymbol{x}^{(p)})\right]\times$$
$$\times\frac{\partial f_i(\boldsymbol{x}^{(p)})}{\partial\boldsymbol{x}}\nabla U(\boldsymbol{x}^{(p)})=0$$

Thus

$$\lambda_p=\frac{\sum\limits_{i=1}^{n}f_i(\boldsymbol{x}^{(p)})\frac{\partial f_i(\boldsymbol{x}^{(p)})}{\partial\boldsymbol{x}}\nabla U(\boldsymbol{x}^{(p)})}{\sum\limits_{i=1}^{n}\left[\frac{\partial f_i(\boldsymbol{x}^{(p)})}{\partial\boldsymbol{x}}\nabla U(\boldsymbol{x}^{(p)})\right]^2}=\frac{(\boldsymbol{f}(\boldsymbol{x}^{(p)}),\ W(\boldsymbol{x}^{(p)})\nabla U(\boldsymbol{x}^{(p)}))}{(W(\boldsymbol{x}^{(p)})\nabla U(\boldsymbol{x}^{(p)}),\ W(\boldsymbol{x}^{(p)})\nabla U(\boldsymbol{x}^{(p)}))}$$

where

$$W(\boldsymbol{x})=\frac{d\boldsymbol{f}}{d\boldsymbol{x}}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n}\\ \cdot & \cdot & \cdots & \cdot\\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n}\end{bmatrix}$$

is the Jacobi matrix of the vector function f.

Furthermore, we have

$$\frac{\partial U}{\partial x_j}=\frac{\partial}{\partial x_j}\left\{\sum_{i=1}^{n}[f_i(\boldsymbol{x})]^2\right\}=2\sum_{i=1}^{n}f_i(\boldsymbol{x})\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}$$

whence

$$\nabla U(\boldsymbol{x})=2\begin{bmatrix}\sum\limits_{i=1}^{n}\frac{\partial f_i(\boldsymbol{x})}{\partial x_1}f_i(\boldsymbol{x})\\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\\ \sum\limits_{i=1}^{n}\frac{\partial f_i(\boldsymbol{x})}{\partial x_n}f_j(\boldsymbol{x})\end{bmatrix}=2W'(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})$$

where $W'(x)$ is the transpose of the Jacobi matrix.
And so, finally,

$$\mu_p=2\lambda_p=\frac{(\boldsymbol{f}^{(p)},\ W_pW'_p\boldsymbol{f}^{(p)})}{(W_pW'_p\boldsymbol{f}^{(p)},\ W_pW'_p\boldsymbol{f}^{(p)}}\tag{5}$$

where, for brevity,

$$\boldsymbol{f}^{(p)}=\boldsymbol{f}(\boldsymbol{x}^{(p)}),\quad W_p=W(\boldsymbol{x}^{(p)})$$

and

$$\boldsymbol{x}^{(p+1)}=\boldsymbol{x}^{(p)}-\mu_pW'_p\boldsymbol{f}^{(p)}\qquad(p=0,\ 1,\ 2,\ \ldots)\tag{6}$$

If we assume that the function $\boldsymbol{f}(\boldsymbol{x})$ is twice continuously differentiable in the neighbourhood of the desired root $\boldsymbol{x}$, then we can obtain more precise formulas for the corrections

$$\Delta\boldsymbol{x}^{(p)}=\boldsymbol{x}^{(p+1)}-\boldsymbol{x}^{(p)}\ \text{(see [7])}$$

Example. Use the method of steepest descent to approximate the roots of the system

$$\left.\begin{aligned}x+x^2-2yz&=0.1,\\ y-y^2+3xz&=-0.2,\\ z+z^2+2xy&=0.3\end{aligned}\right\}$$

located in the neighbourhood of the origin.

Solution. We have

$$\boldsymbol{x}^{(0)}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

Here

$$f=\begin{bmatrix} x+x^2-2yz-0.1 \\ y-y^2+3xz+0.2 \\ z+z^2+2xy-0.3 \end{bmatrix}$$

and

$$W=\begin{bmatrix} 1+2x & -2z & -2y \\ 3z & 1-2y & 3x \\ 2y & 2x & 1+2z \end{bmatrix}$$

Substituting the zeroth approximation, we obtain

$$f^{(0)}=\begin{bmatrix} -0.1 \\ 0.2 \\ -0.3 \end{bmatrix} \quad \text{and} \quad W_0=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}=E$$

Using formulas (5) and (6), we get the first approximation

$$\mu_0=\frac{(f^{(0)},\ f^{(0)})}{(f^{(0)},\ f^{(0)})}=1$$

and

$$x^{(1)}=x^{(0)}-1\cdot Ef^{(0)}=\begin{bmatrix} 0.1 \\ -0.2 \\ 0.3 \end{bmatrix}$$

Analogously, we find the second approximation $x^{(1)}$. We have

$$f^{(1)}=\begin{bmatrix} 0.13 \\ 0.05 \\ 0.05 \end{bmatrix}, \quad W_1=\begin{bmatrix} 1.2 & -0.6 & 0.4 \\ 0.9 & 1.4 & 0.3 \\ -0.4 & 0.2 & 1.6 \end{bmatrix}$$

whence

$$W_1'f^{(1)}=\begin{bmatrix} 0.181 \\ 0.002 \\ 0.147 \end{bmatrix}$$

and

$$W_1W_1'f^{(1)}=\begin{bmatrix} 0.2748 \\ 0.2098 \\ 0.1632 \end{bmatrix}$$

Hence

$$\mu_1=\frac{0.13\cdot 0.2748+0.05\cdot 0.2098+0\cdot 05\cdot 0.1632}{0.2748^2+0.2098^2+0.1632^2}=\frac{0.054374}{0.14619797}=0.3719$$

and

$$x^{(2)}=\begin{bmatrix} 0.1 \\ -0.2 \\ 0.3 \end{bmatrix}-0.37119\cdot\begin{bmatrix} 0.181 \\ 0.002 \\ 0.147 \end{bmatrix}=\begin{bmatrix} 0.0327 \\ -0.2007 \\ 0.2453 \end{bmatrix}$$

To check, we compute the residual

$$f^{(2)} = \begin{bmatrix} 0.032 \\ -0.017 \\ -0.007 \end{bmatrix}$$

13.13 THE METHOD OF STEEPEST DESCENT FOR THE CASE OF A SYSTEM OF LINEAR EQUATIONS

Consider the system of linear equations

$$\left.\begin{aligned} f_1 &\equiv \sum_{j=1}^{n} a_{1j}x_j - b_1 = 0, \\ f_2 &\equiv \sum_{j=1}^{n} a_{2j}x_j - b_2 = 0, \\ &\cdots\cdots\cdots\cdots \\ f_n &\equiv \sum_{j=1}^{n} a_{nj}x_j - b_n = 0 \end{aligned}\right\} \qquad (1)$$

with real matrix $A = [a_{ij}]$ and the column of constant terms

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix}$$

Then

$$\boldsymbol{f} = \mathrm{A}\boldsymbol{x} - \boldsymbol{b}$$

and

$$W = \frac{d\boldsymbol{f}}{d\boldsymbol{x}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} = A$$

Consequently

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \mu_p A' \boldsymbol{r}_p \qquad (2)$$

where $\boldsymbol{r}_p = A\boldsymbol{x}^{(p)} - \boldsymbol{b}$ is the **residual** of the vector $\boldsymbol{x}^{(p)}$ and

$$\mu_p = \frac{(\boldsymbol{r}_p,\ AA'\boldsymbol{r}_p)}{(AA'\boldsymbol{r}_p,\ AA'\boldsymbol{r}_p)} \qquad (p = 0,\ 1,\ 2,\ \dots) \qquad (3)$$

(cf. [5], [6]).

Use of formulas (2) and (3) leads to unwieldy computations. For this reason, one simply uses "descent" instead of "steepest descent" in practical situations, striving towards the minimum of the function

$$U=(A\boldsymbol{x}-\boldsymbol{b},\quad A\boldsymbol{x}-\boldsymbol{b})$$

Here, generally speaking, the number of steps in the process that ensure a given accuracy of the roots of system (1) increases. But it is possible to achieve a situation in which the computation of each step is simpler.

In the general statement of the problem, we assume

$$\boldsymbol{x}^{(p+1)}=\boldsymbol{x}^{(p)}-\lambda_p\boldsymbol{y}^{(p)}\qquad(p=0,\ 1,\ 2,\ \ldots)$$

where $\boldsymbol{y}^{(p)}$ is an arbitrary vector directed to the outside of the level surface $U=\text{const}$ passing through the point $\boldsymbol{x}^{(p)}$, i.e.,

$$(\operatorname{grad} U\ (\boldsymbol{x}^{(p)}),\ \boldsymbol{y}^{(p)})>0$$

We have

$$\boldsymbol{r}_{p+1}=A\boldsymbol{x}^{(p+1)}-\boldsymbol{b}=A\boldsymbol{x}^{(p)}-\boldsymbol{b}-\lambda_pA\boldsymbol{y}^{(p)}=\boldsymbol{r}_p-\lambda_pA\boldsymbol{y}^{(p)}$$

A possible way of determining the scalar factor λ_p proceeds from the requirement [7]

$$(\boldsymbol{r}_{p+1},\ \boldsymbol{y}^{(p)})=(\boldsymbol{r}_p,\ \boldsymbol{y}^{(p)})-\lambda_p\ (A\boldsymbol{y}^{(p)},\ \boldsymbol{y}^{(p)})=0$$

whence

$$\lambda_p=\frac{(\boldsymbol{r}_p,\ \boldsymbol{y}^{(p)})}{(A\boldsymbol{y}^{(p)},\boldsymbol{y}^{(p)})}$$

Depending on the choice of the vector $\boldsymbol{y}^{(p)}$, a variety of computational schemes result. In particular, if the matrix $A=A'$ is positive definite (Sec. 10.15), then, putting $\boldsymbol{y}^{(p)}=\boldsymbol{r}_p$, we have

$$\boldsymbol{x}^{(p+1)}=\boldsymbol{x}^{(p)}-\frac{(\boldsymbol{r}_p,\ \boldsymbol{r}_p)}{(A\boldsymbol{r}_p,\ \boldsymbol{r}_p)}\boldsymbol{r}_p$$

$(p=0, 1, 2, \ldots)$, and $(\operatorname{grad} U\ (\boldsymbol{x}^{(p)}),\boldsymbol{y}^{(p)})=2\,(A\boldsymbol{r}_p,\ \boldsymbol{r}_p)>0$ for $\boldsymbol{r}_p\neq0$.

Example. Use the method of steepest descent to solve the system of equations

$$\left.\begin{aligned}8x_1-x_2-2x_3&=2.3,\\10x_2+x_3+2x_4&=-0.5,\\-x_1+6x_3+2x_4&=-1.2,\\3x_1-x_2+2x_3+12x_4&=3.7,\end{aligned}\right\}\qquad(4)$$

Solution. Since the matrix of the system is dominated by the diagonal elements, for the initial vector $\boldsymbol{x}^{(0)}$ we take the vector whose coordinates are rounded values of the roots of the system:

$$8x_1=2.3,\quad 6x_3=-1.2,$$
$$10x_2=-0.5,\quad 12x_4=3.7,$$

Then, for example,

$$x^{(0)}=\begin{bmatrix}0.3\\-0.05\\-0.2\\0.3\end{bmatrix}$$

Hence

$$r_0=Ax^{(0)}-b=\begin{bmatrix}8&-1&-2&0\\0&10&1&2\\-1&0&6&2\\3&-1&2&12\end{bmatrix}\begin{bmatrix}0.3\\-0.05\\-0.2\\0.3\end{bmatrix}-\begin{bmatrix}2.3\\-0.5\\-1.2\\3.7\end{bmatrix}=\begin{bmatrix}0.55\\0.4\\0.3\\0.45\end{bmatrix}$$

Furthermore

$$A'r_0=\begin{bmatrix}8&0&-1&3\\-1&10&0&-1\\-2&1&6&2\\0&2&2&12\end{bmatrix}\begin{bmatrix}0.55\\0.4\\0.3\\0.45\end{bmatrix}=\begin{bmatrix}5.45\\3.0\\2.0\\6.8\end{bmatrix}$$

and

$$AA'r_0=\begin{bmatrix}8&-1&-2&0\\0&10&1&2\\-1&0&6&2\\3&-1&2&12\end{bmatrix}\begin{bmatrix}5.45\\3.0\\2.0\\6.8\end{bmatrix}=\begin{bmatrix}36.6\\45.6\\20.15\\98.95\end{bmatrix}$$

Using formula (3), we get

$$\mu_0=\frac{(r_0,\ AA'r_0)}{(AA'r_0,\ AA'r_0)}=$$

$$=\frac{0.55\cdot 36.6+0.4\cdot 45.6+0.3\cdot 20.15+0.45\cdot 98.95}{36.6^2+45.6^2+20.15^2+98.95^2}=\frac{88.9425}{13616.0452}=0.006532$$

whence

$$x^{(1)}=x^{(0)}-\mu_0A'r_0=\begin{bmatrix}0.3\\-0.05\\-0.2\\0.3\end{bmatrix}-0.006\,532\begin{bmatrix}5.45\\3.0\\2.0\\6.8\end{bmatrix}=\begin{bmatrix}0.2644\\-0.0696\\-0.2131\\0.2556\end{bmatrix}$$

and

$$r^{(1)}=Ax^{(1)}-b=\begin{bmatrix}0.3109\\0.1020\\0.1684\\-0.1966\end{bmatrix}$$

Subsequent approximations and the corresponding residuals are found in a similar manner:

$$x^{(2)}=\begin{bmatrix}0.2351\\-0.0849\\-0.2147\\0.2863\end{bmatrix},\quad r_2=\begin{bmatrix}0.0956\\0.0087\\0.2493\\0.0967\end{bmatrix},$$

$$x^{(3)}=\begin{bmatrix}0.2296\\-0.0842\\-0.2251\\0.2748\end{bmatrix},\quad r_3=\begin{bmatrix}0.0712\\-0.0280\\0.1692\\-0.0806\end{bmatrix},$$

$$x^{(4)}=\begin{bmatrix}0.2266\\-0.0792\\-0.2379\\0.2875\end{bmatrix},\quad r_4=\begin{bmatrix}0.0680\\0.0354\\0.1211\\0.0334\end{bmatrix},$$

$$x^{(5)}=\begin{bmatrix}0.2228\\-0.0810\\-0.2430\\0.2823\end{bmatrix},\quad r_5=\begin{bmatrix}0.0493\\0.0013\\0.0839\\-0.0493\end{bmatrix}$$

and so on.

It will be seen that the approximation process converges slowly in this case: after the fifth approximation, we are still a good distance away from the exact roots of system (4), which are $x_1=0.2$, $x_2=-0.1$, $x_3=-0.3$, $x_4=0.3$.

*13.14 THE METHOD OF POWER SERIES

Suppose there is a nonlinear system

$$f_k(x_1, x_2, \ldots, x_n)=0 \tag{1}$$

($k=1, 2, \ldots, n$), where the functions f_k are analytic in the neighbourhood of the isolated solution $x^*=(x_1^*, x_2^*, \ldots, x_n^*)$.

Let us consider the more general system [8]

$$F_k(x_1, x_2, \ldots, x_n;\ \lambda)=0 \tag{2}$$

($k=1, 2, \ldots, n$) which depends on a real parameter λ and is such that for $\lambda=0$ the system (2) is solved directly, while for $\lambda=1$ it is identical to system (1), i. e.,

$$F_k(x_1, x_2, \ldots, x_n;\ 1)\equiv f_k(x_1, x_2, \ldots, x_n)$$

($k=1, 2, \ldots, n$). The parameter λ should be introduced so that the dependence of the functions F_k on λ is as simple as possible. For instance, if $x^{(0)}=(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ is a rough approximation to the solution, then we can put

$$\sum_{j=1}^{n}(x_j - x_j^{(0)})\frac{\partial f_k(\boldsymbol{x}^{(0)})}{\partial x_j} + \lambda\left[f_k(\boldsymbol{x}) - \sum_{j=1}^{n}(x_j - x_j^{(0)})\frac{\partial f_k(\boldsymbol{x}^{(0)})}{\partial x_j}\right] = 0$$

$(k = 1, 2, \ldots, n)$, where

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

We will assume that F_k are analytic functions of λ for $|\lambda| \leqslant 1$.

Suppose that for $|\lambda| \leqslant 1$ the system (2) has a simple analytic solution $x_j(\lambda)$ $(j = 1, 2, \ldots, n)$, which for $\lambda = 1$ coincides with x_j^* $(j = 1, 2, \ldots, n)$. Put

$$x_j(0) = x_j^{(0)} \qquad (j = 1, 2, \ldots, n)$$

where $x_j^{(0)}$ $(j = 1, 2, \ldots, n)$ is a known solution of system (2) for $\lambda = 0$. Expanding the functions $x_j(\lambda)$ in a Taylor series at the point $\lambda = 0$, we obtain

$$x_j(\lambda) = x_j(0) + \lambda x_j'(0) + \frac{\lambda^2}{2!}x_j''(0) + \ldots \qquad (j = 1, 2, \ldots, n) \tag{3}$$

To determine the coefficients $x_j'(0)$ differentiate (2) with respect to the parameter λ:

$$\sum_{j=1}^{n}\frac{\partial F_k}{\partial x_j}x_j'(\lambda) + \frac{\partial F_k}{\partial \lambda} = 0 \qquad (k = 1, 2, \ldots, n) \tag{4}$$

Assuming $\boldsymbol{x} = \boldsymbol{x}^{(0)}$ and $\lambda = 0$, we have

$$\sum_{j=1}^{n}\frac{\partial F_k(\boldsymbol{x}^{(0)}; 0)}{\partial x_j}x_j'(0) = -\frac{\partial F_k(\boldsymbol{x}^{(0)}; 0)}{\partial \lambda} \qquad (k = 1, 2, \ldots, n)$$

whence if

$$\det\left[\frac{\partial F_k(\boldsymbol{x}^{(0)}; 0)}{\partial x_j}\right] \neq 0$$

we find $x_j'(0)$.

Furthermore, differentiating (4) with respect to λ, we obtain

$$\sum_{j=1}^{n}\frac{\partial F_k}{\partial x_j}x_j''(\lambda) + \sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\partial^2 F_k}{\partial x_j\,\partial x_l}x_j'(\lambda)\,x_l'(\lambda) +$$

$$+ 2\sum_{j=1}^{n}\frac{\partial^2 F_k}{\partial x_j \partial \lambda}x_j'(\lambda) + \frac{\partial^2 F_k}{\partial \lambda^2} = 0$$

whence, for $\boldsymbol{x}=\boldsymbol{x}^{(0)}$ and $\lambda=0$, we find

$$\sum_{j=1}^{n} \frac{\partial F_k(\boldsymbol{x}^{(0)};\,0)}{\partial x_j} x_j''(0) = -\sum_{j=1}^{n}\sum_{l=1}^{n} \frac{\partial^2 F_k(\boldsymbol{x}^{(0)};\,0)}{\partial x_j \partial x_l} x_j'(0)\, x_l'(0) - $$

$$-2\sum_{j=1}^{n} \frac{\partial^2 F_k(\boldsymbol{x}^{(0)};\,0)}{\partial x_j \partial \lambda} x_j'(0) - \frac{\partial^2 F_k(\boldsymbol{x}^{(0)};\,0)}{\partial \lambda^2} \qquad (k=1,\,2,\,\ldots,\,n) \qquad (5)$$

Since $x_j'(0)$ are known, from system (5) it is possible to determine $x_j''(0)$. The derivatives $x'''(0)$, $x^{\mathrm{IV}}(0)$, ... are computed in a similar manner.

Note that the matrix of coefficients of the higher-order derivatives proves to be the same all the time and to be equal to the Jacobi matrix of the functions $F_1, F_2, \ldots, F_n$ with respect to the variables $x_1, x_2, \ldots, x_n$ for $x_j = x_j^{(0)}$ $(j=1,\,2,\,\ldots,\,n)$ and $\lambda=0$.

Assuming that the series (3) converge for $\lambda=1$, we finally get

$$\boldsymbol{x}_j^* = x_j(1) = x_j(0) + x_j'(0) + \frac{1}{2!} x_j''(0) + \ldots \qquad (j=1,\,2,\,\ldots,\,n) \qquad (6)$$

A defect of the method is the complexity of the computations in the general case of higher-order derivatives. What is more, the convergence of series (6) may not be fast enough.

When using the method, one need not assume the functions $x_j(\lambda)$ $(j=1,\,2,\,\ldots,\,n)$ to be analytic; namely, in place of the Taylor series, one can take advantage of Taylor's formula, terminating the series $x_j(\lambda)$ with some power λ^s and estimating their remainders by familiar formulas (Sec. 3.4).

REFERENCES FOR CHAPTER 13

[1] *L. V. Kantorovich, On Newton's Method,* 1949 (in Russian).

[2] *Alexander Ostrowski, Sur la convergence et l'estimation des erreurs dans quelques procédés de résolution des équations numériques,* 1940.

[3] *James B. Scarborough, Numerical Mathematical Analysis,* 1955, Chapter IX.

[4] *D. A. Ventsel, E. S. Ventsel, Elements of the Theory of Approximate Computations,* 1949, Chapter III, Sec. 8 (in Russian).

[5] *William E. Milne, Numerical Solution of Differential Equations,* 1953, Chapter 9.

[6] *Alston S. Householder, Principles of Numerical Analysis,* 1953, Chapter 3.

[7] *E. But, Numerical Methods,* 1959 (in Russian).

[8] *Edwin F. Beckenbach* (editor), *Modern Mathematics for the Engineer,* First Series, 1956, Chapter 16, Nonlinear Methods by Charles B. Morrey, Jr.

Chapter 14

THE INTERPOLATION OF FUNCTIONS

14.1 FINITE DIFFERENCES OF VARIOUS ORDERS

Suppose

$$y = f(x)$$

is a given function. We denote by $\Delta x = h$ the fixed value of the increment in the argument (*interval* or *spacing*). Then the expression

$$\Delta y \equiv \Delta f(x) = f(x + \Delta x) - f(x) \qquad (1)$$

is called the *first finite difference* of the function y. *Finite differences of higher orders* are defined in similar fashion:

$$\Delta^n y = \Delta(\Delta^{n-1} y) \qquad (n = 2, 3, \ldots)$$

For example,

$$\Delta^2 y = \Delta[f(x + \Delta x) - f(x)] = [f(x + 2\Delta x) - f(x + \Delta x)] -$$
$$- [f(x + \Delta x) - f(x)] = f(x + 2\Delta x) - 2f(x + \Delta x) + f(x)$$

Example. Construct finite differences for the function

$$P(x) = x^3$$

taking the interval as $\Delta x = 1$.

Solution. We have

$$\Delta P(x) = (x + 1)^3 - x^3 = 3x^2 + 3x + 1,$$
$$\Delta^2 P(x) = [3(x + 1)^2 + 3(x + 1) + 1] - (3x^2 + 3x + 1) = 6x + 6,$$
$$\Delta^3 P(x) = [6(x + 1) + 6] - (6x + 6) = 6,$$
$$\Delta^n P(x) = 0 \text{ for } n > 3$$

Note that the finite difference of order three of the function $P(x)$ is constant.

Generally, the following assertion holds true: if

$$P_n(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n$$

is an nth degree polynomial, then $\Delta^n P_n(x)=n!\,a_0h^n=\text{const}$, where $\Delta x=h$.

Indeed, we have

$$\Delta P_n(x)=P_n(x+h)-P_n(x)=a_0\,[(x+h)^n-x^n]+$$
$$+a_1\,[(x+h)^{n-1}-x^{n-1}]+\ldots+a_{n-1}\,[(x+h)-h]$$

Expanding the parentheses by the binomial theorem, we readily see that $\Delta P_n(x)$ is a polynomial of degree $n-1$:

$$\Delta P_n(x)=b_0x^{n-1}+b_1x^{n-2}+\ldots+b_{n-1}$$

where

$$b_0=nha_0$$

Reasoning in a similar manner, we conclude that the second difference $\Delta^2 P_n(x)$ is a polynomial of degree $n-2$:

$$\Delta^2 P_n(x)=c_0x^{n-2}+c_1x^{n-3}+\ldots+c_{n-2}$$

and

$$c_0=(n-1)\,hb_0=n\,(n-1)\,h^2a_0$$

Continuing in this manner successively, we finally establish that

$$\Delta^n P_n(x)=n!a_0h^n=\text{constant}$$

As a corollary we obtain

$$\Delta^s P_n(x)=0 \quad \text{for} \quad s>n$$

The symbol Δ (delta) may be regarded as an *operator* that associates the function $\Delta y=f(x+\Delta x)-f(x)$ (Δx constant) with the function $y=f(x)$. It is easy to verify the basic properties of the operator Δ:

(1) $\Delta\,(u+v)=\Delta u+\Delta v$,

(2) $\Delta\,(Cu)=C\Delta u$ (C constant),

(3) $\Delta^m\,(\Delta^n y)=\Delta^{m+n}y$

where m and n are nonnegative integers, and $\Delta^0 y=y$ by definition.

From the formula (1) we have

$$f(x+\Delta x)=f(x)+\Delta f(x)$$

whence, regarding Δ as a symbolic multiplier, we obtain

$$f(x+\Delta x)=(1+\Delta)f(x) \tag{2}$$

Applying this relation n times in succession, we have

$$f(x+n\Delta x)=(1+\Delta)^n f(x) \tag{3}$$

Taking advantage of the binomial formula,[1] we finally derive

$$f(x+n\Delta x)=\sum_{m=0}^{n} C_n^m \Delta^m f(x) \tag{4}$$

where

$$C_n^m=\frac{n(n-1)\ldots[n-(m-1)]}{m!}$$

is the number of combinations of n elements taken m at a time.

Thus, with the aid of formula (4) the successive values of the function $f(x)$ are expressed in terms of its finite differences of various orders.

Using the identity

$$\Delta=(1+\Delta)-1 \tag{5}$$

and applying the binomial theorem, we obtain

$$\Delta^n f(x)=[(1+\Delta)-1]^n f(x)=(1+\Delta)^n f(x)-C_n^1(1+\Delta)^{n-1}f(x)+$$
$$+C_n^2(1+\Delta)^{n-2}f(x)-\ldots+(-1)^n f(x)$$

whence, by formula (3), we get

$$\Delta^n f(x)=f(x+n\Delta x)-C_n^1 f[x+(n-1)\Delta x]+$$
$$+C_n^2 f[x+(n-2)\Delta x]-\ldots+(-1)^n f(x) \tag{6}$$

Formula (6) expresses the nth-order finite difference of the function $f(x)$ in terms of successive values of the function.

Suppose $f(x)$ has a continuous derivative $f^{(n)}(x)$ on the interval $[x, x+n\Delta x]$. Then the following important formula holds true:

$$\Delta^n f(x)=(\Delta x)^n f^{(n)}(x+\theta n\Delta x) \tag{7}$$

where

$$0<\theta<1$$

---

[1] It is left to the reader to substantiate the legitimacy of using the binomial formula.

The easiest way to prove (7) is by mathematical induction.

Indeed, for $n=1$ we get the mean-value theorem and, hence, formula (7) is true. Now, for $k<n$, we have

$$\Delta^k f(x) = (\Delta x)^k f^{(k)}(x+\theta' k\Delta x)$$

where

$$0<\theta'<1$$

Then

$$\Delta^{k+1} f(x) = \Delta^k [f(x+\Delta x) - f(x)] =$$
$$= (\Delta x)^k [f^{(k)}(x+\Delta x+\theta' k\Delta x) - f^{(k)}(x+\theta' k\Delta x)]$$

Applying tha mean-value theorem to the resulting increment of the derivative $f^{(k)}(x)$, we have

$$\Delta^{k+1} f(x) = (\Delta x)^k \Delta x f^{(k+1)}(x+\theta' k\Delta x+\theta''\Delta x)$$

where $0<\theta''<1$. Assuming

$$\frac{\theta' k+\theta''}{k+1} = \theta \tag{8}$$

we finally get

$$\Delta^{k+1} f(x) = (\Delta x)^{k+1} f^{(k+1)}(x+\theta(k+1)\Delta x)$$

and, obviously,

$$0<\theta<1$$

Thus, the transition from k to $k+1$ is established and, hence, formula (7) is proved.

From formula (7) we have

$$f^{(n)}(x+\theta n\Delta x) = \frac{\Delta^n f(x)}{(\Delta x)^n}$$

Then, passing to the limit as $\Delta x \to 0$ and assuming that the derivative $f^{(n)}(x)$ is continuous, we obtain

$$f^{(n)}(x) = \lim_{\Delta x \to 0} \frac{\Delta^n f(x)}{(\Delta x)^n} \tag{9}$$

Thus, for small Δx, the approximate formula

$$f^{(n)}(x) \approx \frac{\Delta^n f(x)}{(\Delta x)^n} \tag{10}$$

is valid.

14.2 DIFFERENCE TABLE

One often has to consider functions $y=f(x)$ specified by tabular values $y_i=f(x_i)$ for a set of equidistant points x_i $(i=0, 1, 2, \ldots)$, where

$$\Delta x_i = x_{i+1} - x_i = h = \text{constant}$$

The finite differences of the sequence y_i are naturally defined by the relations

$$\begin{aligned} \Delta y_i &= y_{i+1} - y_i, \\ \Delta^2 y_i &= \Delta(\Delta y_i) = \Delta y_{i+1} - \Delta y_i, \\ &\ldots\ldots\ldots\ldots \\ \Delta^n y_i &= \Delta(\Delta^{n-1} y_i) = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i \end{aligned}$$

From the first equation we have

$$y_{i+1} = y_i + \Delta y_i = (1+\Delta)\, y_i$$

whence we derive in succession

$$\begin{aligned} y_{i+2} &= (1+\Delta)\, y_{i+1} = (1+\Delta)^2 y_i, \\ y_{i+3} &= (1+\Delta)\, y_{i+2} = (1+\Delta)^3 y_i, \\ &\ldots\ldots\ldots\ldots\ldots \\ y_{i+n} &= (1+\Delta)^n y_i \end{aligned}$$

Using the binomial theorem, we obtain

$$y_{i+n} = y_i + C_n^1 \Delta y_i + C_n^2 \Delta^2 y_i + \ldots + \Delta^n y_i$$

Conversely, we have

$$\begin{aligned} \Delta^n y_i = [(1+\Delta)-1]^n y_i = (1+\Delta)^n y_i - C_n^1 (1+\Delta)^{n-1} y_i + \\ + C_n^2 (1+\Delta)^{n-2} y_i - \ldots + (-1)^n y_i \end{aligned}$$

or

$$\Delta^n y_i = y_{n+i} - C_n^1 y_{n+i-1} + C_n^2 y_{n+i-2} - \ldots + (-1)^n y_i$$

For example,

$$\begin{aligned} \Delta^2 y_i &= y_{i+2} - 2y_{i+1} + y_i, \\ \Delta^3 y_i &= y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i \end{aligned}$$

and so on. Note that to compute the nth difference $\Delta^n y_i$ one has to know $n+1$ terms $y_i, y_{i+1}, \ldots, y_{i+n}$ of the given sequence.

Finite differences of various orders are conveniently arranged in the form of two types of tables: a *horizontal difference table* (Table 33) or a *diagonal difference table* (Table 34).

TABLE 33
HORIZONTAL DIFFERENCE TABLE

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0	Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$
x_1	y_1	Δy_1	$\Delta^2 y_1$	$\Delta^3 y_1$
x_2	y_2	Δy_2	$\Delta^2 y_2$	$\Delta^3 y_2$
..	..	..	...	....
..	..	..	...	....

TABLE 34
DIAGONAL DIFFERENCE TABLE

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0			
		Δy_0		
x_1	y_1		$\Delta^2 y_0$	
		Δy_1		$\Delta^3 y_0$
x_2	y_2		$\Delta^2 y_1$	
		Δy_2		
x_3	y_3			

Example 1. Form a horizontal difference table for the function

$$y = 2x^3 - 2x^2 + 3x - 1 \tag{1}$$

from the initial value $x_0 = 0$, using $h = 1$ as the interval.

Solution. Assuming $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, we find the corresponding values $y_0 = -1$, $y_1 = 2$, $y_2 = 13$, whence we get

$$\Delta y_0 = y_1 - y_0 = 3,$$
$$\Delta y_1 = y_2 - y_1 = 11,$$
$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = 8$$

These values are entered in the table (Table 35). Since our function is a polynomial of degree three, the third difference is constant (see Sec. 14.1) and equal to

TABLE 35
HORIZONTAL DIFFERENCE TABLE
OF THE CUBIC FUNCTION (1)

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0	—1	3	8	12
1	2	11	20	12
2	13	31	32	12
3	44	63	44	12
4	107	107	56	12
5	214	163	68	12
. . . .	. . . .	. . . .	. . . .	. . . .

$$\Delta^3 y_i = 2 \cdot 3! = 12$$

And so the rest of the table (Table 35) can be filled in by means of summation, using the formulas

$$\begin{aligned} \Delta^2 y_{i+1} &= \Delta^2 y_i + 12 \qquad && (i = 0,\ 1,\ 2,\ \ldots), \\ \Delta y_{i+1} &= \Delta y_i + \Delta^2 y_i && (i = 1,\ 2,\ \ldots), \\ y_{i+1} &= y_i + \Delta y_i && (i = 2,\ 3,\ \ldots) \end{aligned}$$

The stepwise broken line indicates the initial data for compiling the table.

Note. Random errors of the computor may crop up when compiling a difference table. Let us see how the error ε in the value of y_n affects the values of the differences. Compiling the appropriate diagonal difference table, we get Table 36.

From Table 36 it is seen that: (1) if y_n contains an error, then also do the differences

$$\Delta y_{n-1}, \qquad \Delta y_n; \qquad \Delta^2 y_{n-2}, \qquad \Delta^2 y_{n-1},\ \Delta^2 y_n$$

and so on; (2) errors enter the kth differences $\Delta^k y$ with binomial coefficients of alternating sign, namely, the errors have, respectively, the values

$$C_k^0 \varepsilon, \qquad -C_k^1 \varepsilon, \qquad C_k^2 \varepsilon,\ \ldots,\ (-1)^k C_k^k \varepsilon$$

TABLE 36

DIAGONAL DIFFERENCE TABLE OF THE CUBIC FUNCTION (1)

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
……	……	……	……	……	……
……	……	……	……	……	……
x_{n-4}	y_{n-4}				
		Δy_{n-4}			
x_{n-3}	y_{n-3}		$\Delta^2 y_{n-4}$		
		Δy_{n-3}		$\Delta^3 y_{n-4}$	
x_{n-2}	y_{n-2}		$\Delta^2 y_{n-3}$		$\Delta^4 y_{n-4}+\varepsilon$
		Δy_{n-2}		$\Delta^3 y_{n-3}+\varepsilon$	
x_{n-1}	y_{n-1}		$\Delta^2 y_{n-2}+\varepsilon$		$\Delta^4 y_{n-3}-4\varepsilon$
		$\Delta y_{n-1}+\varepsilon$		$\Delta^3 y_{n-2}-3\varepsilon$	
x_n	$y_n+\varepsilon$		$\Delta^2 y_{n-1}-2\varepsilon$		$\Delta^4 y_{n-2}+6\varepsilon$
		$\Delta y_n-\varepsilon$		$\Delta^3 y_{n-1}+3\varepsilon$	
x_{n+1}	y_{n+1}		$\Delta^2 y_n+\varepsilon$		$\Delta^4 y_{n-1}-4\varepsilon$
		Δy_{n+1}		$\Delta^3 y_n-\varepsilon$	
x_{n+2}	y_{n+2}		$\Delta^2 y_{n+1}$		$\Delta^4 y_n+\varepsilon$
		Δy_{n+2}		$\Delta^3 y_{n+1}$	
x_{n+3}	y_{n+3}		$\Delta^2 y_{n+2}$		
		Δy_{n+3}			
x_{n+4}	y_{n+4}				

and, consequently, the absolute value of the maximum error of the kth difference grows rapidly with the number of the difference; (3) for each finite difference $\Delta^k y$, the sum of the errors (with regard for sign) is zero, while the sum of the absolute values of the errors is equal to $|\varepsilon| \cdot 2^k$. Thus, even a slight error in the value of the function leads to considerable errors in high-order differences. Note that in the case of a diagonal difference table the maximum error of the differences $\Delta^k y$ lies in the same horizontal row as the erroneous tabular value y_n, or in the adjacent upper and lower rows.

This *propagation law of the ε-error* in difference tables makes it possible in certain cases to detect and locate an error and also to find its numerical value, thus enabling one to rectify it.

Difference tables are ordinarily compiled to an accuracy of some fixed decimal place. If the function $y=f(x)$ has continuous derivatives up to the mth order, then, given a sufficiently small interval $h=\Delta x$, its differences up to the mth order inclusive vary smoothly, and the mth difference is nearly constant within the limits of the given decimal places. Any violation of this condition in some section of a table generally indicates a computational error (if the function is without singularities).

Having established the maximum deviation of the mth difference from the norm, one can locate the error in the column of values of the function y on the assumption that (1) the error is single and consists in an erroneous computation of one value of the function, and (2) no new errors were made in computing the finite differences. If such an error is detected in a difference table, it can be rectified with the aid of difference values. We show how this is done and for the sake of simplicity confine ourselves to the case of constancy of second or third differences.

Suppose the erroneous tabular value is $y_n+\varepsilon$, where the subscript n has been established and the magnitude of the error ε is unknown.

If the third differences are practically constant, then the second differences form an arithmetic progression, and for this reason the true value of the second difference $\Delta^2 y_{n-1}$ will be equal to the arithmetic mean of three successive erroneous differences:

$$\Delta^2 y_{n-1}=\frac{1}{3}\left[(\Delta^2 y_{n-2}+\varepsilon)+(\Delta^2 y_{n-1}-2\varepsilon)+(\Delta^2 y_n+\varepsilon)\right]$$

since the terms in ε cancel out.

Using the true value of the second difference $\Delta^2 y_{n-1}$ thus found, it is possible to find the magnitude of the error ε; namely, this error is equal to one-half the difference between the corrected and erroneous values of the difference $\Delta^2 y_{n-1}$:

$$\varepsilon=\frac{1}{2}\left[\Delta^2 y_{n-1}-(\Delta^2 y_{n-1}-2\varepsilon)\right]$$

Then the true value of the function y_n itself can be found from the identity

$$y_n=(y_n+\varepsilon)-\varepsilon$$

As a check, compute the differences once again.

Example 2. Correct the error in Table 37.

TABLE 37

DIFFERENCE TABLE CONTAINING AN ERROR

x	y	Δy	$\Delta^2 y$	Error
15	13.260			
		884		
16	14.144		0	
		884		
17	15.028		0	
		884		ε
18	15.912		(—4) 0	
		88(0)4		
19	16.79(2)6		(8) 0	—2 ε
		88(8)4		
20	17.680		(—4) 0	ε
		884		
21	18.564		0	
		884		
22	19.448		0	
		884		
23	20.332			

Solution. Here, the smooth course of the second differences becomes most irregular for $x = 19$. The error affects three rows indicated by the brace. Let us find the arithmetic mean of the second difference for the middle one of the three rows

$$\Delta^2 y_{n-1} = \frac{10^{-3}}{3}(-4+8-4) = 0$$

whence

$$\varepsilon = \frac{1}{2}[0 - 0.008] = -0.004$$

Correcting the tabular value of y for $x = 19$, we get

$$y_n = (y_n + \varepsilon) - \varepsilon = 16.792 - (-0.004) = 16.796$$

After the correction, we get a table with regular variation of the first differences and a constant second difference (the erroneous digits are given in brackets). Bear in mind that this method is good only for correcting isolated computational errors or copying errors. There are special "smoothing" techniques [1] designed to eliminate large numbers of errors that may appear for a variety of reasons, and also to diminish any accumulation of errors resulting from inaccuracies of the computational methods themselves and from rounding off intermediate results to a given number of decimal places.

14.3 GENERALIZED POWER

In the sequel we will need a concept which we will call a *generalized power* (see [1] where the term '*factorial*' is used for this notion).

Definition. The generalized nth power of a number x is a product of n factors, the first of which is equal to x, and each subsequent one is h less than the preceding:

$$x^{[n]}=x(x-h)(x-2h)\dots[x-(n-1)h] \tag{1}$$

where h is some fixed constant.

The exponent of a generalized power is enclosed in square brackets. It is agreed that $x^{[0]}=1$.

For $h=0$, the generalized power (1) coincides with the ordinary power:

$$x^{[n]}=x^n$$

Let us compute the differences for a generalized power, assuming $\Delta x=h$. For the first difference we have

$$\begin{aligned}\Delta x^{[n]}&=(x+h)^{[n]}-x^{[n]}=\\&=(x+h)x\dots[x-(n-2)h]-x(x-h)\dots[x-(n-1)h]=\\&=x(x-h)\dots[x-(n-2)h]\cdot\{(x+h)-[x-(n-1)h]\}=\\&=x(x-h)\dots[x-(n-2)h]\,nh=nhx^{[n-1]}\end{aligned}$$

that is

$$\Delta x^{[n]}=nhx^{[n-1]} \tag{2}$$

Compute the second difference:

$$\Delta^2x^{[n]}=\Delta(\Delta x^{[n]})=\Delta(nhx^{[n-1]})=nh\cdot(n-1)hx^{[n-2]}=nh^2(n-1)x^{[n-2]}$$

Thus

$$\Delta^2x^{[n]}=n(n-1)h^2x^{[n-2]}$$

Using the method of mathematical induction it is easy to prove the general formula

$$\Delta^kx^{[n]}=n(n-1)\dots[n-(k-1)]\,h^kx^{[n-k]}$$

where $k=1, 2, \dots, n$.

It is obvious that

$$\Delta^kx^{[n]}=0 \quad \text{for} \quad k>n$$

From formula (2) there also follows a simple formula for *finite summation*. Let

$$x_0, x_1, x_2, \ldots$$

be equally spaced points with interval h:

$$x_{i+1} - x_i = h \qquad (i = 0, 1, 2, \ldots)$$

Consider the sum

$$S_N = \sum_{i=0}^{N-1} x_i^{[n]}$$

Since, by formula (2), we have

$$x^{[n]} = \frac{\Delta x^{[n+1]}}{h(n+1)}$$

it follows that

$$S_N = \frac{1}{h(n+1)} \sum_{i=0}^{N-1} \Delta x_i^{[n+1]} =$$

$$= \frac{1}{h(n+1)} \{x_1^{[n+1]} - x_0^{[n+1]} + x_2^{[n+1]} - x_1^{[n+1]} + \ldots + x_N^{[n+1]} - x_{N-1}^{[n+1]}\} =$$

$$= \frac{1}{h(n+1)} (x_N^{[n+1]} - x_0^{[n+1]})$$

Thus

$$\sum_{i=0}^{N-1} x_i^{[n]} = \frac{x_N^{[n+1]} - x_0^{[n+1]}}{h(n+1)} \tag{3}$$

Formula (3) is similar to the Newton-Leibniz formula for a positive integral power.

14.4 STATEMENT OF THE PROBLEM OF INTERPOLATION

The simplest problem of interpolation [2] consists in the following. On an interval $[a, b]$ are specified $n+1$ points $x_0, x_1, \ldots, x_n$, called *mesh points (interpolation points)*, and the values of some function $f(x)$ at these points:

$$f(x_0) = y_0, \quad f(x_1) = y_1, \quad \ldots, \quad f(x_n) = y_n \tag{1}$$

It is required to construct a function $F(x)$ *(interpolating function)* belonging to a known class and assuming the same values at the interpolation points as $f(x)$, that is, such that

$$F(x_0) = y_0, \quad F(x_1) = y_1, \quad \ldots, \quad F(x_n) = y_n \tag{2}$$

Geometrically, this means that one has to find a curve $y=F(x)$ of some specific type that passes through the given set of points $M_i(x_i, y_i)$ $(i=0, 1, 2, \dots)$ (Fig. 61).

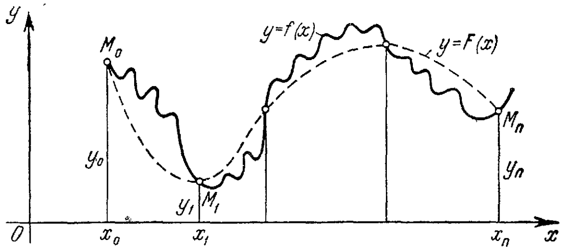


Fig. 61

In such a general statement, the problem can have an infinity of solutions or none at all. However, the problem becomes unambiguous if in place of the arbitrary function $F(x)$ we seek a polynomial $P_n(x)$ of degree not higher than n that satisfies the conditions (2); that is, such that

$$P_n(x_0)=y_0, \qquad P_n(x_1)=y_1, \ \dots, \ P_n(x_n)=y_n$$

The resulting interpolation formula

$$y=F(x)$$

is ordinarily used to approximate the values of the given function $f(x)$ for values of the argument x that differ from the interpolation points. This operation is called *interpolation of the function* $f(x)$. We distinguish between *interpolation in the narrow sense* when $x \in [x_0, x_n]$, that is the value of x is intermediate between x_0 and x_n, and *extrapolation*, when $x \notin [x_0, x_n]$. In the sequel we will use the term *interpolation* to mean both interpolation in the narrow sense and extrapolation.

14.5 NEWTON'S FIRST INTERPOLATION FORMULA

Suppose we have a function $y=f(x)$ and are given the values $y_i=f(x_i)$ for equally spaced values of the independent variable: $x_i=x_0+ih$ $(i=0, 1, 2, \dots, n)$, where h is the *spacing (interval)*. It is required to find a polynomial $P_n(x)$ of degree not higher than n and assuming at points x_i the values

$$P_n(x_i)=y_i \qquad (i=0, 1, \dots, n) \tag{1}$$

Conditions (1) are equivalent to

$$\Delta^m P_n(x_0)=\Delta^m y_0$$

for $m=0, 1, 2, \dots, n$.

Following Newton, we seek the polynomial in the form

$$\begin{aligned}P_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \\ + a_3(x-x_0)(x-x_1)(x-x_2) + \\ + \ldots + a_n(x-x_0)(x-x_1)\ldots(x-x_{n-1})\end{aligned} \tag{2}$$

Using the generalized power, we write expression (1) as

$$\begin{aligned}P_n(x) = a_0 + a_1(x-x_0)^{[1]} + a_2(x-x_0)^{[2]} + \\ + a_3(x-x_0)^{[3]} + \ldots + a_n(x-x_0)^{[n]}\end{aligned} \tag{2'}$$

Our problem consists in determining the coefficients a_i $(i=0, 1, 2, \ldots, n)$ of the polynomial $P_n(x)$. Setting $x = x_0$ in (2′), we obtain

$$P_n(x_0) = y_0 = a_0$$

To find the coefficient a_1, form the first difference

$$\begin{aligned}\Delta P_n(x) = a_1 h + 2a_2(x-x_0)^{[1]}h + \\ + 3a_3(x-x_0)^{[2]}h + \ldots + na_n(x-x_0)^{[n-1]}h\end{aligned}$$

Putting $x = x_0$ in this expression, we get

$$\Delta P_n(x_0) = \Delta y_0 = a_1 h$$

whence

$$a_1 = \frac{\Delta y_{(0)}}{1!h}$$

To determine the coefficient a_2, form the second difference

$$\begin{aligned}\Delta^2 P_n(x) = 2!h^2 a_2 + 2\cdot 3h^2 a_3(x-x_0)^{[1]} + \\ + \ldots + (n-1)nh^2 a_n(x-x_0)^{[n-2]}\end{aligned}$$

Putting $x = x_0$, we get

$$\Delta^2 P_n(x_0) = \Delta^2 y_0 = 2!h^2 a_2$$

whence

$$a_2 = \frac{\Delta^2 y_0}{2!\,h^2}$$

Continuing this process successively, we find that

$$a_i = \frac{\Delta^i y_0}{i!\,h^i} \qquad (i = 0, 1, 2, \ldots, n)$$

where

$$0! = 1 \quad \text{and} \quad \Delta^0 y = y$$

Substituting the values of the coefficients a_i thus found into expression (2′), we get *Newton's interpolation polynomial*

$$P_n(x)=y_0+\frac{\Delta y_0}{1!\,h}(x-x_0)^{[1]}+\frac{\Delta^2 y_0}{2!\,h^2}(x-x_0)^{[2]}+ \\ +\ldots+\frac{\Delta^n y_0}{n!\,h^n}(x-x_0)^{[n]} \qquad (3)$$

It is easy to see that the polynomial (3) fully satisfies the requirements of the problem. Indeed, firstly, the degree of the polynomial $P_n(x)$ does not exceed n, secondly,

$$P_n(x_0)=y_0$$

and

$$P_n(x_k)=y_0+\frac{\Delta y_0}{h}(x_k-x_0)+\frac{\Delta^2 y_0}{2!\,h^2}(x_k-x_0)(x_k-x_1)+\ldots \\ \ldots+\frac{\Delta^k y_0}{k!\,h^k}(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})= \\ =y_0+k\Delta y_0+\frac{k(k-1)}{2!}\Delta^2 y_0+\ldots+\frac{k(k-1)\ldots 1}{k!}\Delta^k y_0= \\ =(1+\Delta)^k y_0=y_k \qquad (k=1,\ 2,\ \ldots,\ n)$$

Note that, as $h\longrightarrow 0$, formula (3) becomes a Taylor polynomial for the function y.

Thus,

$$\lim_{h\to 0}\frac{\Delta^k y_0}{h^k}=\left(\frac{d^k y}{dx^k}\right)_{x=x_0}=y^{(k)}(x_0)$$

Besides, clearly,

$$\lim_{h\to 0}(x-x_0)^{[n]}=(x-x_0)^n$$

whence, as $h\longrightarrow 0$, formula (3) assumes the aspect of the Taylor polynomial:

$$P_n(x)=y(x_0)+y'(x_0)(x-x_0)+\ldots+\frac{y^{(n)}x_0}{n!}(x-x_0)^n$$

For practical use, the Newton interpolation formula (3) is ordinarily written in a modified form. For this purpose we introduce a new variable q via the formula

$$q=\frac{x-x_0}{h}$$

Then

$$\frac{(x-x_0)^{[i]}}{h^i}=\frac{(x-x_0)}{h}\cdot\frac{(x-x_0-h)}{h}\cdot\frac{(x-x_0-2h)}{h}\ldots \\ \ldots\frac{[x-x_0-(i-1)h]}{h}=q(q-1)(q-2)\ldots(q-i+1) \\ (i=1,\ 2,\ \ldots,\ n)$$

Putting these expressions into (3), we get

$$P_n(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots$$
$$\ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0 \qquad (4)$$

where $q = \frac{x-x_0}{h}$ is the *number of steps* needed to reach point x proceeding from point x_0. This is the final form of *Newton's first interpolation formula.*

Formula (4) is conveniently used for interpolating a function $y = f(x)$ **in the neighbourhood of the initial value** x_0, where q is small in absolute value.

If we put $n = 1$ in (4), then we get a formula for *linear interpolation:*

$$P_1(x) = y_0 + q\Delta y_0$$

For $n = 2$ we have the formula for *parabolic interpolation:*

$$P_2(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2}\Delta^2 y_0$$

If we have an unlimited table of values of the function y, then the number n in the interpolation formula (4) may be arbitrary. In practice, the number n in that case is chosen so that the difference $\Delta^n y_i$ is a constant to a specified degree of accuracy. Any tabular value of the argument x can serve as the initial value x_0.

If the table of values of the function is finite, then the number n is bounded, namely: n cannot be greater than the number of values of the function y diminished by unity.

Note that when using Newton's first interpolation formula, it is convenient to use a horizontal difference table, for then the required values of the differences of the function are found in the appropriate horizontal row of the table.

Example 1. Using a spacing of $h = 0.05$, construct Newton's interpolation polynomial on the interval [3.5, 3.6] for the function $y = e^x$ given by the table

x	3.50	3.55	3.60	3.65	3.70
y	33.115	34.813	36.598	38.475	40.447

Solution. Form a difference table (Table 38). Note that, following accepted practice, in the difference columns we do not indicate decimal orders (which are clear from the column of differences of the function). Since the third differences are practically constant, we put $n=3$ in (4). Taking $x_0=3.50$, $y_0=33.115$, we will have

$$P_3(x)=33.115+1.698q+0.087\frac{q(q-1)}{2}+0.005\frac{q(q-1)(q-2)}{6}$$

or

$$P_3(x)=33.115+1.698q+0.0435q(q-1)+0.00083q(q-1)(q-2)$$

where

$$q=\frac{x-3.50}{0.05}=20(x-3.5)$$

TABLE 38
DIFFERENCE TABLE FOR $y=e^x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
3.50	33.115	1698	87	5
3.55	34.813	1785	92	3
3.60	36.598	1877	95	
3.65	38.475	1972		
3.70	40.447			

Example 2. Table 39 contains the values of the *probability integral*

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}\,dx$$

Applying Newton's first interpolation formula, we get, approximately, $\Phi(1.43)$.

Solution. We extend Table 39 by adding differences of the function y up to order three inclusive.

For x_0 we take the tabular value closest to the desired value $x=1.43$; i.e., we put $x_0=1.4$. Since $h=0.1$, then

$$q=\frac{1.43-1.4}{0.1}=0.3$$

TABLE 39

DIFFERENCE TABLE FOR $y = \Phi(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1.0	0.8427	375	—74	10
1.1	0.8802	301	—64	10
1.2	0.9103	237	—54	9
1.3	0.9340	183	—45	9
1.4	0.9523	138	—36	9
1.5	0.9661	102	—27	5
1.6	0.9763	75	—22	6
1.7	0.9838	53	—16	4
1.8	0.9891	37	—12	
1.9	0.9928	25		
2.0	0.9953			

Substituting into (4), we obtain

$$y \approx 0.9523 + 0.3 \cdot 0.0138 + \frac{0.3\,(0.3-1)}{2!}\,(-0.0036) +$$
$$+ \frac{0.3\,(0.3-1)\,(0.3-2)}{3!} \cdot 0.0009 = 0.95686$$

(Tabular value: $\Phi(1.43) = 0.9569$; see Jahnke and Emde's "*Funktionentafeln*".)

In practical situations it is often necessary, for a function given in tabular form, to find an analytic formula representing the tabular values of the function to a certain degree of accuracy. This is called an *empirical formula*, and the problem of constructing it is ambiguous.

When constructing an empirical formula, one has to take into account the general properties of the function. If it is found, from a difference table, that the nth differences of the function are constant for equally spaced values of the argument, then one can take the corresponding first interpolation formula of Newton for the empirical formula.

Example 3. Construct an empirical formula for the function y specified by the table

x	0	1	2	3	4	5
y	5.2	8.0	10.4	12.4	14.0	15.2

Solution. Forming the difference table (Table 40), we see that the second difference is constant.

TABLE 40

FINITE DIFFERENCES OF THE FUNCTION y

x	y	Δy	$\Delta^2 y$
0	5.2	2.8	—0.4
1	8.0	2.4	—0.4
2	10.4	2.0	—0.4
3	12.4	1.6	—0.4
4	14.0	1.2	
5	15.2		

Using Newton's interpolation formula in the form (3), and noting that $h=1$, we have

$$y = 5.2 + 2.8x - \frac{0.4}{2} x(x-1)$$

or

$$y = 5.2 + 3x - 0.2x^2$$

Example 4. Find the sum of the squares

$$S_n = 1^2 + 2^2 + \ldots + n^2$$

of the natural numbers from 1 to n.

Solution. We clearly have

$$\Delta S_n = S_{n+1} - S_n = (n+1)^2$$

whence

$$\Delta^2 S_n = 2n+3, \qquad \Delta^3 S_n = 2$$

and, thus, S_n may be sought as a third-degree polynomial in n. To determine the differences

$$\Delta S_1, \quad \Delta^2 S_1$$

we have to compute three values: S_1, S_2, and S_3. We have

$$S_1 = 1,$$
$$S_2 = S_1 + 2^2 = 1 + 4 = 5,$$
$$S_3 = S_2 + 3^2 = 5 + 9 = 14$$

and from this

$$\Delta S_1 = 5 - 1 = 4,$$
$$\Delta S_2 = 14 - 5 = 9,$$
$$\Delta^3 S_1 = 9 - 4 = 5$$

and

$$\Delta^3 S_1 = 2$$

Using Newton's first interpolation formula and noting that

$$q = \frac{n-1}{1} = n-1$$

we get

$$S_n = 1 + 4(n-1) + \frac{5(n-1)(n-2)}{2} + \frac{2(n-1)(n-2)(n-3)}{6}$$

or

$$S_n = \frac{1}{6} n(n+1)(2n+1)$$

14.6 NEWTON'S SECOND INTERPOLATION FORMULA

Newton's first interpolation formula is inconvenient for interpolating functions near the end of a table. In this case, one ordinarily takes advantage of *Newton's second interpolation formula*, which we now derive.

Suppose we have a set of values of the function

$$y_i = y(x_i) \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n)$$

for equally spaced values of the argument

$$x_i = x_0 + ih$$

We construct an interpolating polynomial of the following form:

$$\begin{aligned} P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \\ + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots \\ \ldots + a_n(x - x_n)(x - x_{n-1}) \ldots (x - x_1) \end{aligned}$$

or, using the generalized power, we obtain

$$\begin{aligned} P_n(x) = a_0 + a_1(x - x_n)^{[1]} + a_2(x - x_{n-1})^{[2]} + \\ + a_3(x - x_{n-2})^{[3]} + \ldots + a_n(x - x_1)^{[n]} \end{aligned} \qquad (1)$$

Our problem is to determine the coefficients $a_0,\ a_1,\ a_2,\ a_3,\ \ldots,\ a_n$ so that the equations

$$P_n(x_i) = y_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n)$$

hold true. To do this, it is necessary and sufficient that

$$\Delta^i P_n(x_{n-i}) = \Delta^i y_{n-i} \qquad (i = 0,\ 1,\ \ldots,\ n) \qquad (2)$$

Setting $x = x_n$ in formula (1), we have

$$P_n(x_n) = y_n = a_0$$

and, hence,

$$a_0 = y_n$$

Now, taking first differences of the left and right members of (1), we have

$$\Delta P_n(x) = a_1 \cdot 1h + a_2 \cdot 2h (x - x_{n-1})^{[1]} + \\ + a_3 \cdot 3h (x - x_{n-2})^{[2]} + \ldots + a_n nh (x - x_1)^{[n-1]}$$

From this, setting $x = x_{n-1}$ and having regard for relation (2), we get

$$\Delta P_n(x_{n-1}) = \Delta y_{n-1} = a_1 h$$

Hence

$$a_1 = \frac{\Delta y_{n-1}}{h}$$

Similarly, forming the second difference of $P_n(x)$, we get

$$\Delta^2 P_n(x) = a_2 2! h^2 + a_3 3 \cdot 2h^2 (x - x_{n-2})^{[1]} + \ldots \\ \ldots + a_n n(n-1) h^2 (x - x_1)^{[n-2]}$$

Assuming $x = x_{n-2}$, we find

$$\Delta^2 P_n(x_{n-2}) = \Delta^2 y_{n-2} = a_2 2! h^2$$

and, thus,

$$a_2 = \frac{\Delta^2 y_{n-2}}{2! h^2}$$

The type of regularity of the coefficients a_i is clear enough. Applying the method of mathematical induction, it can be demonstrated rigorously that

$$a_i = \frac{\Delta^i y_{n-i}}{i! h^i} \qquad (i = 0, 1, 2, \ldots, n) \tag{3}$$

Substituting these values into (1), we finally have

$$P_n(x) = y_n + \frac{\Delta y_{n-1}}{1! h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{2! h^2}(x - x_n)(x - x_{n-1}) + \\ \ldots + \frac{\Delta^3 y_{n-3}}{3! h^3}(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots \\ \ldots + \frac{\Delta^n y_{(0)}}{n! h^n}(x - x_n) \ldots (x - x_1) \tag{4}$$

Formula (4) is called *Newton's second interpolation formula*. We introduce a more convenient form of (4). Let

$$q = \frac{x - x_n}{h}$$

then

$$\frac{x - x_{n-1}}{h} = \frac{x - x_n + h}{h} = q + 1,$$

$$\frac{x - x_{n-2}}{h} = q + 2, \text{ etc.}$$

Substituting these values into (4), we get

$$P_n(x)=y_n+q\Delta y_{n-1}+\frac{q(q+1)}{2!}\Delta^2 y_{n-2}+$$
$$+\frac{q(q+1)(q+2)}{3!}\Delta^3 y_{n-3}+\ldots+\frac{q(q+1)\ldots(q+n-1)}{n!}\Delta^n y_0 \qquad (4')$$

This is the commonly used form of *Newton's second interpolation formula*. For approximate computations of the values of a function y, we put

$$y=P_n(x)$$

Example 1. Given a seven-place log table of values of $y=\log_{10}x$

x	y
1000	3.0000000
1010	3.0043214
1020	3.0086002
1030	3.0128372
1040	3.0170333
1050	3.0211893

Find $\log_{10} 1044$.

Solution. Form the table of differences (Table 41).

TABLE 41

FINITE DIFFERENCES OF THE FUNCTION $y=\log_{10}x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1000	3.0000000	43 214	—426	8
1010	3.0043214	42 788	—418	9
1020	3.0086002	42 370	—409	8
1030	3.0128372	41 961	—401	—
1040	3.0170333	41 560	—	
1050	3.0211893	—		
—	—			

Assume

$$x_n=1050$$

Then

$$q=\frac{x-x_n}{h}=\frac{1044-1050}{10}=-0.6$$

Using the underlined differences, we have, by virtue of (4′),

$$\log_{10} 1044 = 3.0211893 + (-0.6)\cdot 0.0041560 + \\ + \frac{(-0.6)\cdot(-0.6+1)}{2}\cdot 0.0000401 + \\ + \frac{(-0.6)\cdot(-0.6+1)\cdot(-0.6+2)}{6}\cdot 0.0000008 = 3.0187005$$

The result is correct to all the digits written.

Either the first or the second interpolation formula of Newton can be used for extrapolating the function, that is, for finding the values of the function y for values of the argument x lying beyond the range of the table. If $x < x_0$ and x is close to x_0, then it is best to use Newton's first formula; in this case

$$q = \frac{x - x_0}{h} < 0$$

But if $x > x_n$ and x is close to x_n, then it is more convenient to use Newton's second formula; note that

$$q = \frac{x - x_n}{h} > 0$$

Thus, Newton's first interpolation formula is ordinarily used for *forward interpolation* and *backward extrapolation*, while Newton's second interpolation formula is used for *backward interpolation* and *forward extrapolation*.

Note that, generally speaking, extrapolation is less exact than the operation of interpolation in the narrow sense.

Example 2. Having a table of values of the function $y = \sin x$ from $x = 15°$ to $x = 55°$ at $h = 5°$ intervals, find $\sin 14°$ and $\sin 56°$.

TABLE 42

DIFFERENCE TABLE FOR $y = \sin x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
15°	0.2588	832	—26	—6
20°	0.3420	806	—32	—6
25°	0.4226	774	—38	—6
30°	0.5000	736	—44	—5
35°	0.5736	692	—49	—5
40°	0.6428	643	—54	—3
45°	0.7071	589	—57	
50°	0.7660	532		
55°	0.8192			

Solution. Form the table of differences (Table 42). We see that the third differences of the function y are practically constant and so we can confine ourselves to them.

To compute sin 14° we assume

$$x_0 = 15^\circ \quad \text{and} \quad x = 14^\circ$$

whence

$$q = \frac{14^\circ - 15^\circ}{5^\circ} = -0.2$$

Applying Newton's first interpolation formula and using the underlined differences, we get

$$\sin 14^\circ \doteq 0.2588 + (-0.2)\cdot 0.0832 + \frac{(-0.2)(-1.2)}{2!}(-0.0026) + \\ + \frac{(-0.2)(-1.2)(-2.2)}{3!}(-0.0006) = 0.2419$$

Tables give $\sin 14^\circ = 0.24192$.

To find sin 56° we put

$$x_n = 55^\circ \quad \text{and} \quad x = 56^\circ$$

whence

$$q = \frac{56^\circ - 55^\circ}{5^\circ} = 0.2$$

Applying Newton's second interpolation formula and using the twice underlined differences, we get

$$\sin 56^\circ = 0.8192 + 0.2\cdot 0.0532 + \frac{0.2\cdot 1.2}{2!}(-0.0057) + \\ + \frac{0.2\cdot 1.2\cdot 2.2}{3!}(-0.0003) = 0.8291$$

Tables give $\sin 56^\circ = 0.82904$.

14.7 TABLE OF CENTRAL DIFFERENCES

In the construction of Newton's interpolation formulas, only those values of a function were used which lie on one side of the chosen initial value; these formulas are thus of a one-sided nature.

In many cases, very useful are interpolation formulas that contain both preceding and following values of the function with respect to the initial value. The ones most often used are those which contain differences located in a horizontal row (of a diagonal difference table of the given function) corresponding to the initial values x_0 and y_0 or in the rows immediately adjacent to it. These differences Δy_{-1}, Δy_0, $\Delta^2 y_{-1}$, ... are called *central differences*

TABLE 43

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
x_{-4}	y_{-4}						
		Δy_{-4}					
x_{-3}	y_{-3}		$\Delta^2 y_{-4}$				
		Δy_{-3}		$\Delta^3 y_{-4}$			
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$		$\Delta^4 y_{-4}$		
		Δy_{-2}		$\Delta^3 y_{-3}$		$\Delta^5 y_{-4}$	
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$		$\Delta^6 y_{-4}$
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		$\Delta^6 y_{-2}$
		Δy_1		$\Delta^3 y_0$		$\Delta^5 y_{-1}$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$		
		Δy_2		$\Delta^3 y_1$			
x_3	y_3		$\Delta^2 y_2$				
		Δy_3					
x_4	y_4						

(Table 43), where

$$x_i = x_0 + ih \qquad (i = 0, \pm 1, \pm 2, \ldots), \qquad y_i = f(x_i),$$

$$\Delta y_i = y_{i+1} - y_i \qquad \Delta^2 y_i = \Delta y_{i+1} - \Delta y_i, \text{ etc.}$$

The corresponding interpolation formulas are called *central-difference formulas* and include the formulas of Gauss, Stirling, and Bessel [3].

14.8 GAUSSIAN INTERPOLATION FORMULAS

Let us first derive the Gaussian interpolation formulas. Suppose we have $2n+1$ equally spaced points

$$x_{-n}, x_{-(n-1)}, \ldots, x_{-1}, x_0, x_1, \ldots, x_{n-1}, x_n$$

where

$$\Delta x_i = x_{i+1} - x_i = h = \text{constant} \qquad (i = -n, -(n-1), \ldots, n-1)$$

and we know the values of the function $y=f(x)$ at these points:

$$y_i=f(x_i) \qquad (i=0,\ \pm 1,\ \ldots,\ \pm n)$$

It is required to construct a polynomial $P(x)$ of degree not exceeding $2n$ such that

$$P(x_i)=y_i \quad \text{for} \quad i=0,\ \pm 1,\ \ldots,\ \pm n$$

From this condition it follows that

$$\Delta^k P(x_i)=\Delta^k y_i \tag{1}$$

for all corresponding values of i and k.

We will seek this polynomial in the form

$$\begin{gathered} P(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+ \\ +a_3(x-x_{-1})(x-x_0)(x-x_1)+a_4(x-x_{-1})(x-x_0)(x-x_1)(x-x_2)+ \\ +a_5(x-x_{-2})(x-x_{-1})(x-x_0)(x-x_1)(x-x_2)+\ldots \\ \ldots+a_{2n-1}(x-x_{-(n-1)})\ldots(x-x_{-1})(x-x_0)(x-x_1)\ldots \\ \ldots(x-x_{n-1})+a_{2n}(x-x_{-(n-1)})\ldots(x-x_{-1})(x-x_0)(x-x_1)\ldots \\ \ldots(x-x_{n-1})(x-x_n) \end{gathered} \tag{2}$$

Introducing generalized powers, we obtain

$$\begin{gathered} P(x)=a_0+a_1(x-x_0)^{[1]}+a_2(x-x_0)^{[2]}+a_3(x-x_{-1})^{[3]}+ \\ +a_4(x-x_{-1})^{[4]}+\ldots+a_{2n-1}(x-x_{-(n-1)})^{[2n-1]}+ \\ +a_{2n}(x-x_{-(n-1)})^{[2n]} \end{gathered} \tag{3}$$

If, in computing the coefficients a_i $(i=0,\ 1,\ 2,\ \ldots,\ 2n)$, we apply the same technique as in the derivation of the Newton interpolation formulas and take into account (1), we then get successively

$$a_0=y_0, \quad a_1=\frac{\Delta y_0}{1!\,h}, \quad a_2=\frac{\Delta^2 y_{-1}}{2!\,h^2}, \quad a_3=\frac{\Delta^3 y_{-1}}{3!\,h^3},$$

$$a_4=\frac{\Delta^4 y_{-2}}{4!\,h^4}, \quad \ldots, \quad a_{2n-1}=\frac{\Delta^{2n-1}y_{-(n-1)}}{(2n-1)!\,h^{2n-1}}, \quad a_{2n}=\frac{\Delta^{2n}y_{-n}}{(2n)!\,h^{2n}}$$

Furthermore, introducing the variable

$$q=\frac{x-x_0}{h}$$

and making an appropriate change in formula (3), we get *Gauss' first interpolation formula:*

$$\begin{gathered} P(x)=y_0+q\Delta y_0+\frac{q(q-1)}{2!}\Delta^2 y_{-1}+\frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-1}+ \\ +\frac{(q+1)q(q-1)(q-2)}{4!}\Delta^4 y_{-2}+\frac{(q+2)(q+1)q(q-1)(q-2)}{5!}\Delta^5 y_{-2}+\ldots \\ \ldots+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1}y_{-(n-1)}+\frac{(q+n-1)\ldots(q-n)}{(2n)!}\Delta^{2n}y_{-n} \end{gathered} \tag{4}$$

or, more briefly,

$$P(x)=y_0+q\Delta y_0+\frac{q^{[2]}}{2!}\Delta^2 y_{-1}+\frac{(q+1)^{[3]}}{3!}\Delta^3 y_{-1}+\frac{(q+1)^{[4]}}{4!}\Delta^4 y_{-2}+\ldots+$$
$$+\frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\Delta^{2n-1}y_{-(n-1)}+\frac{(q+n-1)^{[2n]}}{(2n)!}\Delta^{2n}y_{-n} \qquad (4')$$

where $x=x_0+qh$ and $q^{[m]}=q(q-1)\ldots[q-(m-1)]$.

The first interpolation formula of Gauss contains the central differences

$$\Delta y_0,\quad \Delta^2 y_{-1},\quad \Delta^3 y_{-1},\quad \Delta^4 y_{-2},\quad \Delta^5 y_{-2},\quad \Delta^6 y_{-3}$$

(see Table 43 where these differences form the lower broken row indicated by the arrows). In a similar manner we can obtain *Gauss' second interpolation formula* which contains the central differences

$$\Delta y_{-1},\quad \Delta^2 y_{-1},\quad \Delta^3 y_{-2},\quad \Delta^4 y_{-2},\quad \Delta^5 y_{-3},\quad \Delta^6 y_{-3},\quad \ldots$$

(in Table 43 these differences form the upper broken row indicated by the arrows).

Gauss' second interpolation formula is of the form

$$P(x)=y_0+q\Delta y_{-1}+\frac{(q+1)q}{2!}\Delta^2 y_{-1}+\frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-2}+$$
$$+\frac{(q+2)(q+1)q(q-1)}{4!}\Delta^4 y_{-2}+\ldots+$$
$$+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1}y_{-n}+$$
$$+\frac{(q+n)(q+n-1)\ldots(q-n+1)}{(2n)!}\Delta^{2n}y_{-n} \qquad (5)$$

or, more compactly,

$$P(x)=y_0+q\Delta y_{-1}+\frac{(q+1)^{[2]}}{2!}\Delta^2 y_{-1}+\frac{(q+1)^{[3]}}{3!}\Delta^3 y_{-2}+\frac{(q+2)^{[4]}}{4!}\Delta^4 y_{-2}+\ldots$$
$$\ldots+\frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\Delta^{2n-1}y_{-n}+\frac{(q+n)^{[2n]}}{(2n)!}\Delta^{2n}y_{-n} \qquad (5')$$

where

$$x=x_0+qh$$

14.9 STIRLING'S INTERPOLATION FORMULA

Taking the arithmetic mean of the first and second interpolation formulas of Gauss (4) and (5) of Sec. 14.8, we get *Stirling's formula*

$$P(x)=y_0+q\cdot\frac{\Delta y_{-1}+\Delta y_0}{2}+\frac{q^2}{2}\Delta^2 y_{-1}+\frac{q(q^2-1^2)}{3!}\cdot\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2}+$$
$$+\frac{q^2(q^2-1^2)}{4!}\Delta^4 y_{-2}+\frac{q(q^2-1^2)(q^2-2^2)}{5!}\cdot\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2}+$$
$$+\frac{q^2(q^2-1^2)(q^2-2^2)}{6!}\Delta^6 y_{-3}+\ldots$$

$$\ldots+\frac{q(q^2-1^2)(q^2-2^2)(q^2-3^2)\ldots[q^2-(n-1)^2]}{(2n-1)!}\cdot\frac{\Delta^{2n-1}y_{-n}+\Delta^{2n-1}y_{-(n-1)}}{2}+$$
$$+\frac{q^2(q^2-1^2)(q^2-2^2)\ldots[q^2-(n-1)^2]}{(2n)!}\Delta^{2n}y_{-n}$$

where

$$q=\frac{x-x_0}{h}$$

It is easy to see that

$$P(x_i)=y_i \text{ for } i=0, \pm 1, \ldots, \pm n$$

14.10 BESSEL'S INTERPOLATION FORMULA

In addition to *Stirling's formula*, frequent use is made of *Bessel's formula*. To derive this formula we take advantage of Gauss' second interpolation formula (5) (see Sec. 14.8).

We take $2n+1$ equally spaced points

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_0,\ \ldots,\ x_{n-1},\ x_n,\ x_{n+1}$$

with spacing h: let

$$y_i=f(x_i) \qquad (i=-n, \ldots, n+1)$$

be the given values of the function $y=f(x)$.

If we take $x=x_0$ and $y=y_0$ for the initial values, then, using the points x_k $(k=0, \pm 1, \ldots, \pm n)$, we have

$$P(x)=y_0+q\Delta y_{-1}+\frac{(q+1)q}{2!}\Delta^2 y_{-1}+\frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-2}+$$
$$+\frac{(q+2)(q+1)q(q-1)}{4!}\Delta^4 y_{-2}+\ldots+$$
$$+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1}y_{-n}+\frac{(q+n)(q+n-1)\ldots(q-n+1)}{(2n)!}\Delta^{2n}y_{-n} \quad (1)$$

Now take $x=x_1$ and $y=y_1$ for the initial values and use the points x_{1+k} $(k=0, \pm 1, \ldots, \pm n)$. Then

$$\frac{x-x_1}{h}=\frac{x-x_0-h}{h}=q-1$$

and, correspondingly, the indices of all the differences in the right member of (1) will increase by unity. Replacing q by $q-1$ in the right member of (1) and increasing the indices of all differences by 1, we get an auxiliary interpolation formula:

$$P(x)=y_1+(q-1)\Delta y_0+\frac{q(q-1)}{2!}\Delta^2 y_0+$$
$$+\frac{q(q-1)(q-2)}{3!}\Delta^3 y_{-1}+\frac{(q+1)q(q-1)(q-2)}{4!}\Delta^4 y_{-1}+$$

$$+\frac{(q+1)q(q-1)(q-2)(q-3)}{5!}\Delta^5 y_{-2}+\dots$$

$$\dots+\frac{(q+n-2)\dots(q-n)}{(2n-1)!}\Delta^{2n-1}y_{-(n-1)}+\frac{(q+n-1)\dots(q-n)}{(2n)!}\Delta^{2n}y_{-(n-1)} \quad (2)$$

Taking the arithmetic mean of (2) and (4) of Sec. 14.8, we get (after some simple manipulations) the *Bessel interpolation formula*

$$P(x)=\frac{y_0+y_1}{2}+\left(q-\frac{1}{2}\right)\Delta y_0+\frac{q(q-1)}{2}\cdot\frac{\Delta^2 y_{-1}+\Delta^2 y_0}{2}+$$

$$+\frac{\left(q-\frac{1}{2}\right)q(q-1)}{3!}\Delta^3 y_{-1}+\frac{q(q-1)(q+1)(q-2)}{4!}\cdot\frac{\Delta^4 y_{-2}+\Delta^4 y_{-1}}{2}+$$

$$+\frac{\left(q-\frac{1}{2}\right)q(q-1)(q+1)(q-2)}{5!}\Delta^5 y_{-2}+$$

$$+\frac{q(q-1)(q+1)(q-2)(q+2)(q-3)}{6!}\cdot\frac{\Delta^6 y_{-3}+\Delta^6 y_{-2}}{2}+\dots$$

$$\dots+\frac{q(q-1)(q+1)(q-2)(q+2)\dots(q-n)(q+n-1)}{(2n)!}\cdot$$

$$\cdot\frac{\Delta^{2n}y_{-n}+\Delta^{2n}y_{-n+1}}{2}+$$

$$\frac{\left(q-\frac{1}{2}\right)q(q-1)(q+1)(q-2)(q+2)\dots(q-n)(q+n--1)}{(2n+1)!}\Delta^{2n+1}y_{-n} \quad (3)$$

where

$$q=\frac{x-x_0}{h}$$

The Bessel interpolation formula (3), as follows from its derivation, is a polynomial that coincides with the given function $y=f(x)$ in $2n+1$ points

$$x_{-n},\ x_{-(n-1)},\ \dots,\ x_n,\ x_{n+1}$$

in the particular case of $n=1$, we have, neglecting the difference $\Delta^3 y_{-1}$, *Bessel's formula for parabolic interpolation:*

$$P(x)=\frac{y_0+y_0+\Delta y_0}{2}+\left(q-\frac{1}{2}\right)\Delta y_0+\frac{q(q-1)}{2}\cdot\frac{\Delta y_0-\Delta y_{-1}+\Delta y_1-\Delta y_0}{2}$$

or

$$P(x)=y_0+q\,\Delta y_0-q_1(\Delta y_1-\Delta y_{-1})$$

where

$$q_1=\frac{q(1-q)}{4}$$

In Bessel's formula, all terms containing odd differences have the multiplier $q-\frac{1}{2}$; therefore, when $q=\frac{1}{2}$, formula (3) is substantially simplified:

$$P\left(\frac{x_0+x_1}{2}\right)=\frac{y_0+y_1}{2}-\frac{1}{8}\,\frac{\Delta^2y_{-1}+\Delta^2y_0}{2}+\\ +\frac{3}{128}\,\frac{\Delta^4y_{-2}+\Delta^4y_{-1}}{2}-\frac{5}{1024}\,\frac{\Delta^6y_{-3}+\Delta^6y_{-2}}{2}+\ldots\\ \ldots+(-1)^n\,\frac{[1\cdot3\cdot5\ \ldots\ (2n-1)]^2}{2^{2n}\,(2n)!}\,\frac{\Delta^{2n}y_{-n}+\Delta^{2n}y_{-n+1}}{2}$$

This special case of the Bessel formula is called the *formula for interpolating to halves*. If in the Bessel formula (3) we change the variable using the formula $q-\frac{1}{2}=p$, then it takes on a more symmetric aspect:

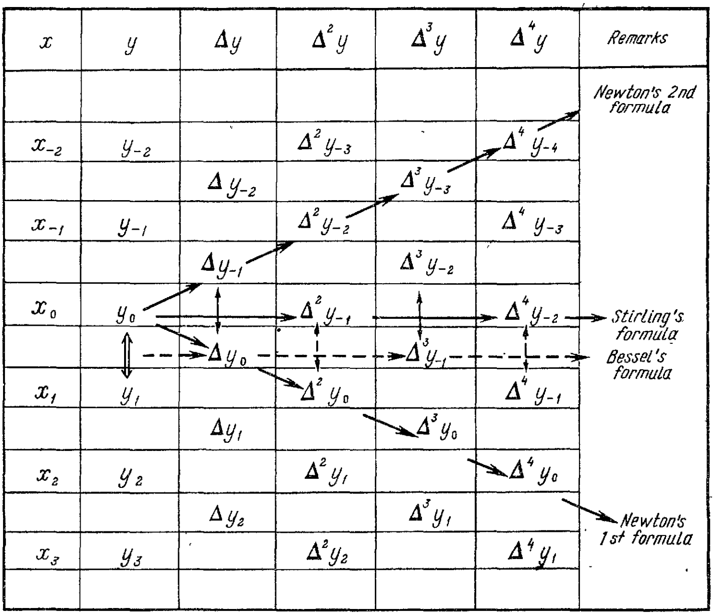
$$P(x)=\frac{y_0+y_1}{2}+p\,\Delta y_0+\frac{\left(p^2-\frac{1}{4}\right)}{2}\cdot\frac{\Delta^2y_{-1}+\Delta^2y_0}{2}+\\ +\frac{p\left(p^2-\frac{1}{4}\right)}{3!}\,\Delta^3y_{-1}+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)}{4!}\cdot\frac{\Delta^4y_{-2}+\Delta^4y_{-1}}{2}+\\ +\frac{p\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)}{5!}\,\Delta^5y_{-2}+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\left(p^2-\frac{25}{4}\right)}{6!}\times\\ \times\frac{\Delta^6y_{-3}+\Delta^6y_{-2}}{2}+\ldots+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n-1)^2}{4}\right]}{(2n)!}\times\\ \times\frac{\Delta^{2n}y_{-n}+\Delta^{2n}y_{-n+1}}{2}+\frac{p\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n-1)^2}{4}\right]}{(2n+1)!}\times\\ \times\Delta^{2n+1}y_{-n+1}\qquad(3')$$

where $p=\frac{1}{h}\left(x-\frac{x_0+x_1}{2}\right)$.

14.11 GENERAL DESCRIPTION OF INTERPOLATION FORMULAS WITH CONSTANT INTERVAL

In a general characterization of interpolation formulas, note the following: when constructing the Newton interpolation formulas, either the first or the last interpolation point is chosen as the initial value x_0; for central formulas of interpolation, a midpoint is taken for the initial value. The scheme given below (Table 44) illustrates the differences used in the basic interpolation formulas; for the purpose of convenience in surveying the table, the numbering of the indices has been changed in Newton's second interpolation formula.

TABLE 44



x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	Remarks
						Newton's 2nd formula
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$		$\Delta^4 y_{-4}$	
		Δy_{-2}		$\Delta^3 y_{-3}$		
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$	
		Δy_{-1}		$\Delta^3 y_{-2}$		
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$	Stirling's formula
		Δy_0		$\Delta^3 y_{-1}$		Bessel's formula
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$	
		Δy_1		$\Delta^3 y_0$		
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$	
		Δy_2		$\Delta^3 y_1$		Newton's 1st formula
x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$	

A more detailed examination of the interpolation formulas shows that it is advisable to use Stirling's formula for $|q| \leqslant 0.25$ and Bessel's formula for $0.25 \leqslant q \leqslant 0.75$. The first and second Newton interpolation formulas are used to advantage when the interpolation is performed at the beginning or, respectively, at the end of a table and the needed central differences are lacking [4].

Example 1. The values of the probability integral [3]

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}\, dx$$

are given in Table 45. Find Φ (0.5437).

Solution. We supplement Table 45 with finite differences of the given function $y = \Phi(x)$. Taking $x_0 = 0.54$ and $x = 0.5437$, we have

$$q = \frac{x - x_0}{h} = \frac{0.5437 - 0.54}{0.01} = 0.37$$

Since $\frac{1}{4} < q < \frac{3}{4}$, we use Bessel's formula (3′) to get

$$p = q - \frac{1}{2} = 0.37 - 0.50 = -0.13$$

TABLE 45

TABLE OF DIFFERENCES OF THE FUNCTION $y = \Phi(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0.51	0.5292437	86550	—896	—7
0.52	0.5378987	85654	—903	—7
0.53	0.5464641	84751	—910	—7
0.54	0.5549392	83841	—917	—6
0.55	0.5633233	82924	—923	
0.56	0.5716157	82001		
0.57	0.5798158			

whence, using the underlined differences, we obtain

$$\Phi(0.5437) = \frac{0.5549392+0.5633233}{2} + (-0.13)\,0.0083841 +$$

$$+\frac{0.0169-0.25}{2}\cdot\frac{-0.0000910-0.0000917}{2}+$$

$$+\frac{-0.13\,(0.0169-0.25)}{6}\cdot(-0.0000007)=$$

$$=0.55913125-0.00108993+0.00001065=0.5580520$$

Example 2. Having Table 46 of the values of the complete elliptic integral

$$K(\alpha)=\int\limits_0^{\frac{\pi}{2}}\frac{dx}{\sqrt{1-\sin^2\alpha\,\sin^2 x}}$$

find $K(78°30')$.

Solution. Put $x_0 = 78°$, $h = 1°$, $x = 78°30'$, whence $q = 0.5$. If we take advantage of Bessel's formula for interpolating to halves, then, confining ourselves to fifth differences, we have

$$K(78°30') = 2.97857 + 0.5\cdot 8316\cdot 10^{-5} -$$

$$-0.125\cdot\frac{715+850}{2}\cdot 10^{-5} + 0.023437\cdot\frac{32+40}{2}\cdot 10^{-5} =$$

$$= 2.97857 - 0.04158 - 0.000978 + 0.000008 = 3.019180$$

TABLE 46

VALUES OF THE COMPLETE ELLIPTIC INTEGRAL $K(\alpha)$

α	$K(\alpha)$	ΔK	$\Delta^2 K$	$\Delta^3 K$	$\Delta^4 K$	$\Delta^5 K$	$\Delta^6 K$
75°	2.76806						
		6461					
76°	2.83267		528				
		6989		84			
77°	2.90256		612		19		
		7601		103		13	
78°	2.97857		715		32		—5
		8316		135		8	
79°	3.06173		850		40		18
		9166		175		26	
80°	3.15339		1025		66		—1
		10191		241		25	
81°	3.25530		1266		91		43
		11457		332		68	
82°	3.36987		1598		159		
		13055		491			
83°	3.50042		2089				
		15144					
84°	3.65186						

Now apply Stirling's formula by way of a comparison:

$$K(78°30') = 2.97857 + 0.5\,\frac{7601+8316}{2}\cdot 10^{-5} +$$

$$+0.125715\cdot 10^{-5} - 0.0625\cdot\frac{103+135}{2}\cdot 10^{-5} - 0.0078\cdot 32\cdot 10^{-5} +$$

$$+0.0117\cdot\frac{13+8}{2}\cdot 10^{-5} = 2.97857 + 0.039792 +$$

$$+0.000894 - 0.000074 - 0.000002 + 0.000001 = 3.019181$$

14.12 LAGRANGE'S INTERPOLATION FORMULA

The interpolation formulas derived in the preceding sections are only suitable for equally spaced points. A more general formula, called *Lagrange's interpolation formula*, is used for arbitrarily specified points.

Let there be given, on an interval $[a, b]$, $n+1$ distinct values of the argument: $x_0, x_1, x_2, \ldots, x_n$ and let the corresponding values of the function $y=f(x)$ be known:

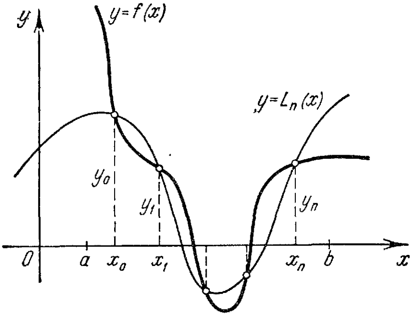
$$f(x_0)=y_0, \quad f(x_1)=y_1, \ \ldots, \ f(x_n)=y_n$$

It is required to construct a polynomial $L_n(x)$ of degree not exceeding n having at the specified points $x_0, x_1, \ldots, x_n$ the same values as the function $f(x)$, that is, such that

$$L_n(x_i)=y_i \qquad (i=0, 1, 2, \ldots, n)$$

(Fig. 62a).

Fig. 62a

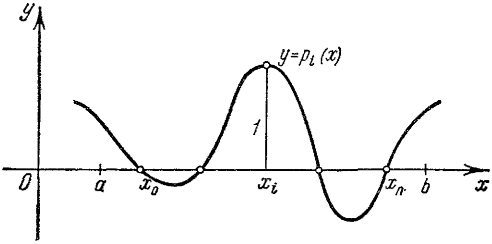


Let us first solve a particular problem: we construct a polynomial $p_i(x)$ such that

$$p_i(x_j)=0 \quad \text{for} \quad j \neq i \quad \text{and} \quad p_i(x_i)=1$$

(Fig. 62b).

Fig. 62b



These conditions can be written compactly thus:

$$p_i(x_j)=\delta_{ij}=\begin{cases} 1 & \text{if} \quad j=i, \\ 0 & \text{if} \quad j \neq i \end{cases} \tag{1}$$

where δ_{ij} is the *Kronecker delta*.

Since the desired polynomial vanishes at n points $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, it has the form

$$p_i(x)=C_i(x-x_0)(x-x_1)\ \ldots\ (x-x_{i-1})(x-x_{i+1})\ \ldots\ (x-x_n) \tag{2}$$

where C_i is a constant coefficient. Setting $x=x_i$ in formula (2) and noting that $p_i(x_i)=1$, we get

$$C_i(x_i-x_0)(x_i-x_1)\ \ldots\ (x_i-x_{i-1})(x_i-x_{i+1})\ \ldots\ (x_i-x_n)=1$$

whence

$$C_i=\frac{1}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}$$

Putting this value in formula (2), we have

$$p_i(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \tag{3}$$

Let us now take up the solution of the general problem: to find a polynomial $L_n(x)$ that satisfies the above-indicated conditions $L_n(x_i)=y_i$.

This polynomial is of the form

$$L_n(x)=\sum_{i=0}^{n} p_i(x)\,y_i \tag{4}$$

Indeed, firstly, it is clear that the degree of the polynomial $L_n(x)$ thus constructed is not higher than n and, secondly, by virtue of Condition (1) we have

$$L_n(x_j)=\sum_{i=0}^{n} p_i(x_j)\,y_i=p_j(x_j)\,y_j=y_j \qquad (j=0,\ 1,\ \ldots,\ n)$$

Substituting the value $p_i(x)$ from (3) into (4), we obtain

$$L_n(x)=\sum_{i=0}^{n} y_i\,\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \tag{5}$$

This is *Lagrange's interpolation formula.*

We will prove the *uniqueness* of the Lagrange polynomial. Assume the contrary.

Let $\tilde{L}_n(x)$ be a polynomial distinct from $L_n(x)$ of degree not exceeding n and such that

$$\tilde{L}_n(x_i)=y_i \qquad (i=0,\ 1,\ \ldots,\ n)$$

Then the polynomial

$$Q_n(x)=\tilde{L}_n(x)-L_n(x)$$

whose degree clearly does not exceed n, vanishes at $n+1$ points $x_0, x_1, x_2, \ldots, x_n$; thus

$$Q_n(x) \equiv 0$$

Hence

$$\tilde{L}_n(x) \equiv L_n(x)$$

From this it follows, for one thing, that if the points are equally spaced, then the Lagrange interpolation polynomial coincides with the corresponding Newton interpolation polynomial.

It will be noted that, in general, all the above-constructed interpolation formulas are obtained from the Lagrange interpolation formula for an appropriate choice of points.

The Lagrange formula (5) may be written compactly if we introduce the following designation:

$$\Pi_{n+2}(x) = (x-x_0)(x-x_1)\ldots(x-x_n) \tag{6}$$

Differentiating this product with respect to x, we get

$$\Pi'_{n+1}(x) = \sum_{j=0}^{n} (x-x_0)(x-x_1)\ \ldots\ (x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)$$

Putting $x = x_i$ $(i = 0, 1, 2, \ldots, n)$, we have

$$\Pi'_{n+1}(x_i) = (x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_0) \tag{7}$$

Putting (6) and (7) into formula (5), we obtain

$$L_n(x) = \Pi_{n+1}(x)\sum_{j=0}^{n}\frac{y_i}{\Pi'_{n+1}(x_i)(x-x_i)} \tag{5'}$$

It is worth noting that the Lagrange formula (unlike the earlier interpolation formulas) contains y_i explicitly, which is sometimes very important.

Let us consider two special cases of Lagrange's interpolation polynomial.

For $n = 1$ we have two points and the Lagrange formula is then the equation of a straight line $y = L_1(x)$ passing through the two given points:

$$y = \frac{x-b}{a-b}y_0 + \frac{x-a}{b-a}y_1$$

where a, b are the abscissas of these points.

For $n = 2$ we get the equation of the parabola $y = L_2(x)$ passing through three points:

$$y = \frac{(x-b)(x-c)}{(a-b)(a-c)}y_0 + \frac{(x-a)(x-c)}{(b-a)(b-c)}y_1 + \frac{(x-a)(x-b)}{(c-a)(c-b)}y_2$$

where a, b, c are the abscissas of the given points.

Example 1. Construct the Lagrange interpolation polynomial for the function $y=\sin \pi x$, choosing the points

$$x_0=0, \quad x_1=\frac{1}{6}, \quad x_2=\frac{1}{2}$$

Solution. Compute the corresponding values of the function:

$$y_0=0, \quad y_1=\sin\frac{\pi}{6}=\frac{1}{2}, \quad y_2=\sin\frac{\pi}{2}=1$$

Applying formula (5), we get

$$L_2(x)=\frac{\left(x-\frac{1}{6}\right)\left(x-\frac{1}{2}\right)}{\left(-\frac{1}{6}\right)\left(-\frac{1}{2}\right)}\cdot 0+\frac{x\left(x-\frac{1}{2}\right)}{\frac{1}{6}\left(\frac{1}{6}-\frac{1}{2}\right)}\cdot\frac{1}{2}+\frac{x\left(x-\frac{1}{6}\right)}{\frac{1}{2}\left(\frac{1}{2}-\frac{1}{6}\right)}\cdot 1$$

or

$$L_2(x)=\frac{7}{2}x-3x^2$$

Example 2. Given a table of the values of the function $y=f(x)$ [3]:

x	y
321.0	2.50651
322.8	2.50893
324.2	2.51081
325.0	2.51188

Compute the value of $f(323.5)$.

Solution. Put $x=323.5$, $n=3$. Then by formula (5) we have

$$\begin{aligned} f(323.5)&=\frac{(323.5-322.8)(323.5-324.2)(323.5-325.0)}{(321-322.8)(321-324.2)(321-325)}\cdot 2.50651+\\ &+\frac{(323.5-321)(323.5-324.2)(323.5-325)}{(322.8-321)(322.8-324.2)(322.8-325)}\cdot 2.50893+\\ &+\frac{(323.5-321)(323.5-322.8)(323.5-325)}{(324.2-321)(324.2-322.8)(324.2-325)}\cdot 2.51081+\\ &+\frac{(323.5-321)(323.5-322.8)(323.5-324.2)}{(325-321)(325-322.8)(325-324.2)}\cdot 2.51188=\\ &=-0.07996+1.18794+1.83897-0.43708=2.50987\end{aligned}$$

*14.13 COMPUTING LAGRANGIAN COEFFICIENTS

We give here a scheme for simplified computation of the coefficients of y_i ($i=0, 1, 2, \ldots, n$) in the Lagrange formula, the so-called *Lagrangian coefficients*

$$L_i^{(n)}(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \qquad (1)$$

or, in a more compact notation,

$$L_i^{(n)}(x)=\frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)} \qquad (2)$$

where

$$\Pi_{n+1}(x)=(x-x_0)\ldots(x-x_n)$$

The Lagrange formula then has the form

$$L_n(x)=\sum_{i=0}^{n} L_i^{(n)}(x)\,y_i$$

It is to be noted that the form of the Lagrangian coefficients is invariant under an integral linear substitution $x=at+b$ (a, b constants and a not zero). Indeed, putting

$$x=at+b,\quad x_j=at_j+b \qquad (j=0,\ 1,\ \ldots,\ n)$$

in formula (1) we get, after cancelling a^n from numerator and denominator,

$$L_i^{(n)}(t)=\frac{(t-t_0)(t-t_1)\ldots(t-t_{i-1})(t-t_{i+1})\ldots(t-t_n)}{(t_i-t_0)(t_i-t_1)\ldots(t_i-t_{i-1})(t_i-t_{i+1})\ldots(t_i-t_n)} \qquad (3)$$

or

$$L_i^{(n)}=\frac{\Pi_{n+1}(t)}{(t-t_i)\,\Pi'_{n+1}(t_i)} \qquad (3')$$

where

$$\Pi_{n+1}(t)=(t-t_0)(t-t_1)\ldots(t-t_n)$$

which completes the proof.

Lagrangian coefficients are conveniently computed (especially by computing machines) by the scheme given below. First enter the differences as follows:

$$\begin{array}{ccccc}
\underline{x-x_0} & x_0-x_1 & x_0-x_2 & \ldots & x_0-x_n \\
x_1-x_0 & \underline{x-x_1} & x_1-x_2 & \ldots & x_1-x_n \\
x_2-x_0 & x_2-x_1 & \underline{x-x_2} & \ldots & x_2-x_n \\
\cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\
\cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\
x_n-x_0 & x_n-x_1 & x_n-x_2 & \ldots & \underline{x-x_n}
\end{array} \qquad (*)$$

Denote the product of the elements of the first row by D_0, of the second row by D_1, and so on. Then the product of the elements of the principal diagonal (these elements are underlined in the accompanying scheme) will obviously be $\Pi_{n+1}(x)$, whence it follows that

$$L_i^{(n)}(x)=\frac{\Pi_{n+1}(x)}{D_i} \qquad (i=0,\ 1,\ \ldots,\ n) \qquad (4)$$

Hence

$$L_n(x) = \Pi_{n+1}(x)\sum_{i=0}^{n}\frac{y_i}{D_i} \tag{5}$$

The Lagrangian coefficients can be reduced to a simpler form in the case of equally spaced points.

Setting

$$x = x_0 + th$$

we have

$$t_0 = 0, \quad t_1 = 1, \ \ldots, \ t_n = n$$

whence

$$\Pi_{n+1}(t) = t(t-1)(t-2)\ldots(t-n)$$

and

$$\Pi'_{n+1}(i) = (-1)^{n-1}\,i!\,(n-i)!$$

Substituting these expressions in formula (3′), we get

$$L_i^{(n)}(t) = \frac{1}{n!}\Pi_{n+1}(t)\cdot\frac{(-1)^{n-i}C_n^i}{t-i} \qquad (i = 0,\ 1,\ \ldots,\ n) \tag{6}$$

where

$$C_n^i = \frac{n!}{i!\,(n-i)!}$$

whence

$$L_n(x) = \frac{1}{n!}\Pi_{n+1}(t)\sum_{i=0}^{n}(-1)^{n-i}\frac{C_n^i}{t-i}\,y_i \tag{7}$$

where

$$t = \frac{x - x_0}{h}$$

In the case of a constant interval h, the problem of interpolation is further simplified by the fact that tables of Lagrangian coefficients are available (see [5]) so that actually all the computations reduce to multiplying the tabulated coefficients by the appropriate values of the function y_i followed by a summation.

Example 1. For the function $y = y(x)$ we have the table

x	0.05	0.15	0.20	0.25	0.35	0.40	0.50	0,55
y	0.9512	0.8607	0.8187	0.7788	0.7047	0.6703	0.6065	0.5769
t	1	3	4	5	7	8	10	11

Find $y(0.45)$.

Solution. To simplify computations we set

$$x = 0.05t$$

Then the values of the new variable t corresponding to the interpolation points will be 1, 3, 4, 5, 7, 8, 10, 11. We have to find the value of y for $x = 0.45$; that is, for $t = 9$. Putting $t = t_i$ $(i = 0, 1, \ldots, 7)$, we arrange the computations as given in the scheme in Table 47.

TABLE 47

COMPUTATIONAL SCHEME FOR LAGRANGIAN COEFFICIENTS

i	$t_i - t_j$ $(j \neq i)$								D_i	y_i	$\frac{y_i}{D_i}$
0	$\boxed{8}$	-2	-3	-4	-6	-7	-9	-10	-725760	0.9512	$-0.0131 \cdot 10^{-4}$
1	2	$\boxed{6}$	-1	-2	-4	-5	-7	-8	26880	0.8607	$0.3202 \cdot 10^{-4}$
2	3	1	$\boxed{5}$	-1	-3	-4	-6	-7	-7560	0.8187	$-1.0829 \cdot 10^{-4}$
3	4	2	1	$\boxed{4}$	-2	-3	-5	-6	5760	0.7788	$1.3520 \cdot 10^{-4}$
4	6	4	3	2	$\boxed{2}$	-1	-3	-4	-3456	0.7047	$-2.0390 \cdot 10^{-4}$
5	7	5	4	3	1	$\boxed{1}$	-2	-3	2520	0.6703	$2.6530 \cdot 10^{-4}$
6	9	7	6	5	3	2	$\boxed{-1}$	-1	11340	0.6065	$0.5348 \cdot 10^{-4}$
7	10	8	7	6	4	3	1	$\boxed{-2}$	-80640	0.5769	$-0.0715 \cdot 10^{-4}$
	$\Pi(9) = 3840$										$S = 1.6535 \cdot 10^{-4}$

whence

$$y(0.45) = \Pi(9) \sum_{i=0}^{i=7} \frac{y_i}{D_i} = \Pi(9) \cdot S = 3840 \cdot 1.6535 \cdot 10^{-4} = \underline{0.6349}$$

Example 2. The function $y = \cos x$ is given by the table [5]

x	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7
y	0.283662185	0.377977743	0.468516671	0.554374336	0.634692876	0.708669774	0.775565879	0.834712785
t	0	1	2	3	4	5	6	7

Find $\cos 5.347$.

Solution. Make a change of variable by the formula

$$x = 0.1t + 5$$

Then the values of the variable t corresponding to the interpolation points will be 0, 1, 2, 3, 4, 5, 6, 7 and the desired value $x = 5.347$ will become $t = 3.47$. Noting that the points $t_i = i$ ($i = 0, 1, \ldots, 7$) are equally spaced, the computations can be carried out by the indicated scheme (see Table 48).

TABLE 48

COMPUTATIONAL SCHEME OF LAGRANGIAN COEFFICIENTS FOR EQUALLY SPACED POINTS

i	x_i	y_i	$t-i$	$(-1)^{7-i}C_7^i$	$(-1)^{7-i}C_7^i \frac{y_i}{t-i}$
0	5.0	0.283662185	3.47	—1	—0.08174702
1	5.1	0.377977743	2.47	7	1.07119198
2	5.2	0.468516671	1.47	—21	—6.69309530
3	5.3	0.554374336	0.47	35	41.28319523
4	5.4	0.634692876	—0.53	—35	41.91368048
5	5.5	0.708669774	—1.53	21	—9.72684003
6	5.6	0.775565879	—2.53	—7	2.14583444
7	5.7	0.834712785	—3.53	1	—0.23646254
		$\Pi = 42.8848749$			$S = 69.67575724$

From Table 48 we have

$$\Pi(3.47) = \prod_{i=0}^{7} (3.47 - i) = 42.8848749$$

and

$$S = \sum_{i=0}^{7} (-1)^{7-i} C_7^i \frac{y_i}{3.47 - i} = 69.67575724$$

On the basis of formula (7) we get

$$\cos 5.347 = \frac{1}{7!} \cdot \Pi(3.47) \cdot S = 0.592864312$$

14.14 ERROR ESTIMATE OF LAGRANGE'S INTERPOLATION FORMULA

In Sec 14.12, for the function $y = f(x)$ we constructed the Lagrange interpolation polynomial $L_n(x)$, which assumes at the points $x_0, x_1, \ldots, x_n$ the given values

$$y_0 = f(x_0), \ y_1 = f(x_1), \ \ldots, \ y_n = f(x_n)$$

The question arises as to how close the constructed polynomial approaches the function $f(x)$ at other points, that is, as to how great is the remainder term

$$R_n(x) = f(x) - L_n(x)$$

To determine this degree of approximation we impose additional restrictions on the function $y = f(x)$, namely, we will assume that in the range under consideration, $a \leqslant x \leqslant b$, which contains the points of interpolation, the function $f(x)$ has all derivatives $f'(x)$, $f''(x)$, $\ldots$, $f^{(n+1)}(x)$ up to the $(n+1)$th order inclusive.

We introduce an auxiliary function

$$u(x) = f(x) - L_n(x) - k\Pi_{n+1}(x) \tag{1}$$

where

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$$

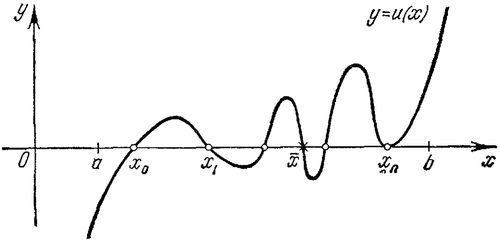
and k is a constant coefficient which will be chosen below.

The function $u(x)$ obviously has $n+1$ roots at the points

$$x_0,\ x_1,\ \ldots,\ x_n$$

Now choose the coefficient k so that $u(x)$ has an $(n+2)$th root at any (fixed) point $\bar{x}$ of the interval $[a, b]$, which point does not coincide with the points of interpolation (Fig. 63). For this it suffices to put

Fig. 63



$$f(\bar{x}) - L_n(\bar{x}) - k\Pi_{n+1}(\bar{x}) = 0$$

whence, since $\Pi_{n+1}(\bar{x}) \neq 0$,

$$k = \frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})} \tag{2}$$

For this value of the factor k, the function $u(x)$ has $n+2$ roots in the interval $[a, b]$ and will vanish at the endpoints of each of the intervals

$$[x_0,\ x_1],\ [x_1,\ x_2],\ \ldots,\ [x_i,\ \bar{x}],\ [\bar{x},\ x_{i+1}],\ \ldots,\ [x_{n-1},\ x_n]$$

Applying Rolle's theorem to each of these intervals, we see that the derivative $u'(x)$ has at least $n+1$ roots in the interval $[a, b]$. Applying Rolle's theorem to the derivative $u'(x)$, we see that the second derivative $u''(x)$ vanishes no less than n times on the interval $[a, b]$.

Continuing this reasoning, we conclude that on the interval $[a, b]$ under consideration, the derivative $u^{(n+1)}(x)$ has at least one zero, which we denote by ξ; thus, $u^{(n+1)}(\xi)=0$.

From formula (1), since

$$L_n^{(n+1)}(x)=0 \quad \text{and} \quad \Pi_{n+1}^{(n+1)}(x)=(n+1)\,!,$$

we have

$$u^{(n+1)}(x)=f^{(n+1)}(x)-k(n+1)\,!$$

For $x=\xi$ we obtain

$$0=f^{(n+1)}(\xi)-k(n+1)\,!$$

whence

$$k=\frac{f^{(n+1)}(\xi)}{(n+1)\,!} \tag{3}$$

Comparing the right members of (2) and (3), we have

$$\frac{f(\bar{x})-L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}$$

Thus

$$f(\bar{x})-L_n(\bar{x})=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\Pi_{n+1}(\bar{x}) \tag{4}$$

Since $\bar{x}$ is arbitrary, formula (4) may be written thus

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\Pi_{n+1}(x) \tag{5}$$

where ξ depends on x and lies inside the interval $[a, b]$.

Note that (5) is valid for all points of $[a, b]$ including the points of interpolation.

Denoting

$$M_{n+1}=\max_{a\leqslant x\leqslant b}|f^{(n+1)}(x)|$$

we obtain the following estimate for the absolute error in the Lagrange interpolation formula:

$$|R_n(x)|=|f(x)-L_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}|\Pi_{n+1}(x)| \tag{6}$$

where

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\dots(x-x_n) \tag{6'}$$

Example. To what degree of accuracy can we calculate $\sqrt{115}$ by means of Lagrange's interpolation formula for the function $y=\sqrt{x}$ if we choose the interpolation points $x_0=100$, $x_1=121$, $x_2=144$?

Solution. We have

$$y'=\frac{1}{2}x^{-\frac{1}{2}},\quad y''=-\frac{1}{4}x^{-\frac{3}{2}},\quad y'''=\frac{3}{8}x^{-\frac{5}{2}}$$

whence

$$M_3=\max|y'''|=\frac{3}{8}\cdot\frac{1}{\sqrt{100^5}}=\frac{3}{8}\cdot 10^{-5}\quad\text{for}\quad 100\leqslant x\leqslant 144.$$

On the basis of formula (6) we obtain

$$|R_2|\leqslant\frac{3}{8}\cdot 10^{-5}\cdot\frac{1}{3!}|(115-100)(115-121)(115-144)|=$$

$$=\frac{1}{16}\cdot 10^{-5}\cdot 15\cdot 6\cdot 29\approx 1.6\cdot 10^{-3}$$

14.15 ERROR ESTIMATES OF NEWTON'S INTERPOLATION FORMULAS

If the points $x_0, x_1, \ldots, x_n$ are equally spaced, and

$$x_{i+1}-x_i=h\qquad(i=0, 1, 2, \ldots, n-1)$$

then, putting

$$q=\frac{x-x_0}{h}$$

we obtain, on the basis of (5) of Sec. 14.14, the *remainder term of Newton's first interpolation formula:*

$$R_n(x)=h^{n+1}\cdot\frac{q(q-1)\ldots(q-n)}{(n+1)!}f^{(n+1)}(\xi)\tag{1}$$

where ξ is a value intermediate between the abscissas $x_0, x_1, \ldots, x_n$ and the point x at hand. Note that for the case of interpolation in the narrow sense $\xi\in[x_0, x_n]$; for extrapolation it is possible that $\xi\notin[x_0, x_n]$.

Similarly, putting

$$q=\frac{x-x_n}{h}$$

in formula (5) of Sec. 14.14, we get the remainder term of Newton's second interpolation formula:

$$R_n(x)=h^{n+1}\cdot\frac{q(q+1)\ldots(q+n)}{(n+1)!}f^{(n+1)}(\xi)\tag{2}$$

where ξ is some value intermediate between the abscissas $x_0, x_1, \ldots, x_n$ and the point x.

Ordinarily, in practical computations, Newton's interpolation formula terminates with terms containing differences which, to within the limits of the specified accuracy, may be considered constant.

Assuming that $\Delta^{n+1}y$ are nearly constant for the function $y=f(x)$ and h is sufficiently small, and taking into account that

$$f^{(n+1)}(x)=\lim_{h\to 0}\frac{\Delta^{n+1}y}{h^{n+1}}$$

we can approximately set

$$f^{(n+1)}(\xi)\approx\frac{\Delta^{n+1}y_0}{h^{n+1}}$$

In this case the remainder term of Newton's first interpolation formula is equal to

$$R_n(x)\approx\frac{q(q-1)\ldots(q-n)}{(n+1)!}\Delta^{n+1}y_0$$

Under these conditions, we get for the remainder term of Newton's second interpolation formula the expression

$$R_n(x)\approx\frac{q(q+1)\ldots(q+n)}{(n+1)!}\Delta^{n+1}y_n$$

Example 1. Five-place log tables give the logarithms of integers from $x=1000$ to $x=10{,}000$ to within a limiting error of $\frac{1}{2}\cdot 10^{-5}$. Is linear interpolation possible to the same degree of accuracy?

Solution. Setting

$$y=\log_{10}x$$

we have

$$y'=\frac{M}{x}\quad\text{and}\quad y''=-\frac{M}{x^2}$$

where $M=0.43$, whence

$$M_2=\max|y''|<\frac{0.5}{10^6}=\frac{1}{2}\cdot 10^{-6}$$

From formula (1) we get, for $n=2$ and $h=1$, an estimate of the error of linear interpolation:

$$|R_1(x)|\leqslant\frac{|q(q-1)|}{2!}M_2\leqslant\frac{q(1-q)}{2}\cdot\frac{1}{2}\cdot 10^{-6}$$

Since for $0\leqslant q\leqslant 1$ we have

$$q(1-q)=\frac{1}{4}-\left(\frac{1}{2}-q\right)^2\leqslant\frac{1}{4}$$

we finally get

$$|R_1(x)| \leqslant \frac{\frac{1}{4}}{2} \cdot \frac{1}{2} \cdot 10^{-6} < 10^{-7}$$

Consequently, linear interpolation is quite admissible.

Example 2. Estimate the error resulting from approximating the function $f(x) = \sin x$ by the fifth degree interpolation polynomial $P_5(x)$ which coincides with the given function for the values $x = 0°, 5°, 10°, 15°, 20°, 25°$.

Solution. Here $f^{(6)}(x) = -\sin x$, and so $|f^{(6)}(x)| \leqslant 1$. On the basis of formula (1) we have

$$|\sin x - P_5(x)| \leqslant \frac{1}{6!} \left| x \left(x - \frac{\pi}{36}\right)\left(x - \frac{\pi}{18}\right)\left(x - \frac{\pi}{12}\right)\left(x - \frac{\pi}{9}\right)\left(x - \frac{5\pi}{36}\right)\right|$$

For example, for $x = 12°30' = \text{arc } 0.21816$ we obtain

$$|\sin x - P_5(x)| < 2.2 \cdot 10^{-9}$$

14.16 ERROR ESTIMATES OF THE CENTRAL INTERPOLATION FORMULAS

We give without proof the remainder terms for the Stirling and Bessel formulas [3].

(a) *The remainder term of Stirling's interpolation formula.* If $2n$ is the order of the largest difference used in a table and $x \in [x_0 - nh, x_0 + nh]$, then

$$R_n(x) = \frac{h^{2n+1} f^{(2n+1)}(\xi)}{(2n+1)!} q(q^2 - 1^2)(q^2 - 2^2)(q^2 - 3^2) \ldots (q^2 - n^2)$$

where

$$q = \frac{x - x_0}{h} \quad \text{and} \quad \xi \in [x_0 - nh, x_0 + nh]$$

But if the analytic expression of the function $f(x)$ is not known, then for small h we assume

$$R_n(x) \approx \frac{\Delta^{2n+1} y_{-n-1} + \Delta^{2n+1} y_{-n}}{2(2n+1)!} q(q^2 - 1^2)(q^2 - 2^2) \ldots (q^2 - n^2)$$

(b) *The remainder term of Bessel's interpolation formula.* If $2n+1$ is the order of the largest difference used in the table and $x \in [x_0 - nh, x_0 + (n+1)h]$, then

$$R_n(x) = \frac{h^{2n+2}}{(2n+2)!} f^{(2n+2)}(\xi) q(q^2 - 1^2)(q^2 - 2^2) \ldots (q^2 - n^2) \times$$
$$\times [q - (n+1)]$$

where

$$q=\frac{x-x_0}{h} \quad \text{and} \quad \xi \in [x_0-nh,\ x_0+(n+1)h]$$

However, if the function $f(x)$ is tabulated and the interval h is small, then we assume

$$R_n(x) \approx \frac{\Delta^{2n+2}y_{-n-1}+\Delta^{2n+2}y_{-n}}{2(2n+2)!}\,q(q^2-1^2)(q^2-2^2)\times\ldots$$
$$\ldots\times(q^2-n^2)[q-(n+1)]$$

In particular, for $q=\frac{1}{2}$ we obtain the error for *interpolating to halves:*

$$R_n=\frac{h^{2n+2}f^{(2n+2)}(\xi)}{(2n+2)!}(-1)^{n+1}\frac{[1\cdot3\cdot5\ \ldots\ (2n+1)]^2}{2^{2n+2}}$$

or

$$R_n \approx \frac{\Delta^{2n+2}y_{-n-1}+\Delta^{2n+2}y_{-n}}{2(2n+2)!}(-1)^{n+1}\frac{[1\cdot3\cdot5\ \ldots\ (2n+1)]^2}{2^{2n+2}}$$

If we put

$$q=p+\frac{1}{2}$$

then the formula for the remainder term of Bessel's formula takes the form

$$R_n(x)=\frac{h^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi)\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n+1)^2}{4}\right]$$

14.17 ON THE BEST CHOICE OF INTERPOLATION POINTS

In analyzing formula (5) of Sec. 14.14, we see that the error $R_n(x)$ in Lagrange's formula is, to within a numerical constant, a product of two factors, one of which, $f^{(n+1)}(\xi)$, depends on the properties of the function $f(x)$ and is not amenable to regulation, while the magnitude of the other, $\Pi_{n+1}(x)$, is determined exclusively by the choice of the interpolation points (abscissas).

If the abscissas x_i are not arranged suitably, the upper bound of the modulus of the error $R_n(x)$ [see (6), Sec. 14.14] may be very large. For example, if we bunch the abscissas x_i near one end of the interval $[a, b]$, then $R_n(x)$ will, generally speaking, for $l=b-a>1$, be great at the points x close to the other end of the interval. The question thus arises of an optimal choice of the interpolation points x_i (for a given number of points n) so that the portion of the error that depends on us—the polynomial $\Pi_{n+1}(x)$—has the least in modulus maximal value on the interval $[a, b]$, or, to put it briefly, "deviates least of all from zero on $[a, b]$". This problem was solved by the Russian mathematician P. L. Chebyshev [2], [6] who proved that the best choice of abscissas (interpolation points) in the indicated meaning is given by the formula

$$x_i = \frac{b+a}{2} + \frac{b-a}{2}\xi_i$$

where

$$\xi_i = -\cos\frac{2i+1}{2n+2}\pi \quad (i = 0, 1, 2, \ldots, n)$$

are zeros of the so-called Chebyshev polynomial $T_{n+1}(x)$. We then have

$$|\Pi_{n+1}(x)| \leqslant 2\left(\frac{b-a}{4}\right)^{n+1}$$

It is interesting to note that these points are not equally spaced but bunch up near the endpoints of the interval. Even for such a choice of points, one cannot, in the general case, guarantee that the absolute value of the error will be arbitrarily small for sufficiently large n.

We will now remark generally on the determination of errors of the interpolation formulas. If the maximal differences are practically constant, then the result of interpolation in the narrow sense is ordinarily correct to as many decimal places as given in the tabulated data, and so it is not necessary to estimate the errors. When using Lagrange's interpolation formula there is no possibility of following the course of finite differences; if possible one should therefore estimate the remainder term.

If the function $f(x)$ is tabulated and the analytic expression is not known, then, strictly speaking, it is impossible to estimate the error of the interpolation polynomial. True enough, since for a given polynomial it is theoretically possible to construct an infinity of distinct functions that coincide with the polynomial in the given set of points, at intermediate points the deviation of the interpolation polynomial from the function may be arbitrarily large. However, if the nature of the function is such that its graph is a smooth curve, then it is possible to determine approximately the errors of the interpolation polynomials to a high degree of confidence on the basis of the values of higher-order differences by the above-indicated formulas.

14.18 DIVIDED DIFFERENCES

When constructing difference tables, we have assumed that the values of the argument of the function are **equally spaced**, that is, that they have a *constant interval*. However, tables for **unequally**

spaced values of the argument (tables with variable interval) are also used. For example, empirical data are of this nature. For tables with unequally spaced values of the argument (variable interval), the concept of finite differences is generalized to so-called *divided differences.*

Suppose we have a tabulated function $y=f(x)$ and x_0, x_1, x_2, ... are the values of the argument and y_0, y_1, y_2, ... are the corresponding values of the function, where the differences

$$x_i=x_{i+1}-x_i \neq 0 \qquad (i=0,\ 1,\ \ldots)$$

are unequal.

The ratios

$$[x_i,\ x_{i+1}]=\frac{y_{i+1}-y_i}{x_{i+1}-x_i}$$

$(i=0,\ 1,\ 2,\ \ldots)$ are called *divided differences of the first order.* For example,

$$[x_0,\ x_1]=\frac{y_1-y_0}{x_1-x_0}, \qquad [x_1,\ x_2]=\frac{y_2-y_1}{x_2-x_1}, \quad \text{etc.}$$

Similarly, we define *second-order divided differences*

$$[x_i,\ x_{i+1},\ x_{i+2}]=\frac{[x_{i+1},\ x_{i+2}]-[x_i,\ x_{i+1}]}{x_{i+2}-x_i}$$

$(i=0,\ 1,\ 2,\ \ldots)$. For example,

$$[x_0,\ x_1,\ x_2]=\frac{[x_1,\ x_2]-[x_0,\ x_1]}{x_2-x_0}$$

and so forth.

Generally, *nth order divided differences* are obtained from $(n-1)$th order divided differences by means of the recurrence relation

$$[x_i,\ x_{i+1},\ \ldots,\ x_{i+n}]=\frac{[x_{i+1},\ \ldots,x_{i+n}]-[x_i,\ \ldots,x_{i+n-1}]}{x_{i+n}-x_i}$$
$$(n=1,\ 2,\ \ldots;\quad i=0,\ 1,\ 2,\ \ldots) \tag{1}$$

Note that divided differences remain unchanged under a permutation of the elements; that is, they are *symmetric functions* of their arguments. For instance,

$$[x_0,\ x_1]=\frac{y_1-y_0}{x_1-x_0}=\frac{y_0-y_1}{x_0-x_1}=[x_1,\ x_0], \quad \text{etc.}$$

Divided differences are ordinarily arranged in tables of the type shown below (Table 49).

TABLE 49

TABLE OF DIVIDED DIFFERENCES

x	y	Divided differences			
		1st	2nd	3rd	4th
x_0	y_0				
		$[x_0, x_1]$			
x_1	y_1		$[x_0, x_1, x_2]$		
		$[x_1, x_2]$		$[x_0, x_1, x_2, x_3]$	
x_2	y_2		$[x_1, x_2, x_3]$		$[x_0, x_1, x_2, x_3, x_4]$
		$[x_2, x_3]$		$[x_1, x_2, x_3, x_4]$	
x_3	y_3		$[x_2, x_3, x_4]$		
		$[x_3, x_4]$			
x_4	y_4				

Example. Form the divided differences for a function specified by the following table:

x	0	0.2	0.3	0.4	0.7	0.9
y	132.651	148.877	157.464	166.375	195.112	216.000

Solution. Successively applying formula (1), we have

$$[x_0,\ x_1]=\frac{148.877-132.651}{0.2-0}\doteq 81.13,$$

$$[x_1,\ x_2]=\frac{157.464-148.877}{0.3-0.2}=85.87,$$

$$[x_0,\ x_1,\ x_2]=\frac{85.87-81.13}{0.3-0}=15.8$$

and so on. The results are tabulated in Table 50.

14.19 NEWTON'S INTERPOLATION FORMULA FOR UNEQUALLY SPACED VALUES OF THE ARGUMENT

Using the concept of divided differences, we can represent Lagrange's interpolation formula in a form similar to Newton's first interpolation formula. First let us prove a lemma of interest in itself.

Lemma. *If* $y=P(x)$ *is an* n*th degree polynomial, then its divided difference of order* $(n+1)$ *is identically zero: that is,*

$$[x,\ x_0,\ x_1,\ \ldots,\ x_n]\equiv 0$$

for any set of distinct numbers $x,\ x_0,\ x_1,\ \ldots,\ x_n$.

TABLE 50

DIVIDED DIFFERENCES OF THE FUNCTION y

x	y	1st	2nd	3rd	4th
0	132.651				
		81.13			
0.2	140.877		15.8		
		85.87		1	
0.3	157.464		16.2		0
		89.11		1	
0.4	166.375		16.7		0
		95.79		1	
0.7	195.112		17.3		
		104.44			
0.9	216.000				

Indeed, if $P(x)$ is a polynomial of degree n, then

$$[x,\ x_0]=\frac{P(x)-P(x_0)}{x-x_0}\equiv P(x,\ x_0)$$

is a polynomial of degree $(n-1)$ in x. Furthermore,

$$[x,\ x_0,\ x_1]=\frac{P(x,\ x_0)-P(x_0,\ x_1)}{x-x_1}\equiv P(x,\ x_0,\ x_1)$$

is a polynomial of degree $(n-2)$ in x. Indeed, the function $P(x,\ x_0)-P(x_0,\ x_1)=P(x,\ x_0)-P(x_1,\ x_0)$ has the root $x=x_1$ and, hence, on the basis of the remainder theorem, the polynomial $P(x,\ x_0)-P(x_0,\ x_1)$ is exactly divisible by the binomial $x-x_1$. By similar reasoning we see that

$$[x,\ x_0,\ \ldots,\ x_{n-1}]\equiv P(x,\ x_0,\ \ldots,\ x_{n-1})$$

is a polynomial of degree zero, that is,

$$P(x,\ x_0,\ \ldots,\ x_{n-1})=C$$

whence

$$[x,\ x_0,\ \ldots,\ x_n]=\frac{C-C}{x-x_n}\equiv 0$$

Now suppose $P(x)$ is a Lagrange polynomial of degree n such that

$$P(x_i)=f(x_i)=y_i \tag{1}$$

$(i=0,\ 1,\ \ldots,\ n)$ where $y=f(x)$ is the given function. Denote by $P(x,\ x_0)$, $P(x,\ x_0,\ x_1)$, $\ldots$, $P(x,\ x_0,\ \ldots,\ x_n)$ the successive divided differences of $P(x)$. We have

$$\left.\begin{aligned} &P(x_0, x_1) = [x_0, x_1], \\ &P(x_0, x_1, x_2) = [x_0, x_1, x_2], \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &P(x_0, x_1, \ldots, x_n) = [x_0, x_1, \ldots, x_n] \end{aligned}\right\} \tag{2}$$

Besides, on the basis of the lemma

$$P(x, x_0, \ldots, x_n) = 0 \tag{3}$$

By definition we obtain

$$\frac{P(x) - P(x_0)}{x - x_0} = P(x, x_0) \tag{4}$$

whence

$$P(x) = P(x_0) + P(x, x_0)(x - x_0) \tag{5}$$

From the definition of divided differences it follows that

$$P(x, x_0, \ldots, x_m) = \frac{P(x, x_0, \ldots, x_{m-1}) - P(x_0, \ldots, x_m)}{x - x_m}$$

whence

$$\begin{aligned} P(x, x_0, \ldots, x_{m-1}) = P(x_0, \ldots, x_m) + \\ + (x - x_m) P(x, x_0, \ldots, x_m) \quad (m = 1, 2, \ldots, n) \end{aligned} \tag{6}$$

Using (6), we obtain successively from (5)

$$\begin{aligned} P(x) &= P(x_0) + P(x, x_0)(x - x_0) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &+ P(x, x_0, x_1)(x - x_0)(x - x_1) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &+ P(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots \\ \ldots &+ P(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1)\ldots(x - x_{n-1}) + \\ &+ P(x, x_0, \ldots, x_n)(x - x_0)(x - x_1)\ldots(x - x_n) \end{aligned}$$

or, taking into account the equations (2) and (3), we finally obtain *Newton's interpolation formula for unequally spaced values of the argument*

$$\begin{aligned} P(x) = y_0 + [x_0, x_1](x - x_0) + [x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots \\ \ldots + [x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-1}) \end{aligned} \tag{7}$$

The error in formula (7) is, as usual, equal to

$$R(x) = f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\ldots(x - x_n) \tag{8}$$

where ξ is an intermediate value between the points $x_0, x_1, \ldots, x_n$ and x.

Example. Form the interpolation polynomial for the function $y=f(x)$ given by the table

x	0	2.5069	5.0154	7.52270
y	0.3989423	0.3988169	0.3984408	0.3978138

Using this polynomial we find $f(3.7608)$.

Solution. We find the divided differences of the function y (Table 51).

TABLE 51

DIVIDED DIFFERENCES OF THE FUNCTION y

x	y	1st	2nd	3rd
0	0.3989423			
		—500		
2.5069	0.3988169		—199	
		—1499		0
5.0154	0.3984408		—199	
		—2496		
7.5270	0.3978138			

Using formula (7), we find

$$y=0.3989423-0.0000500x-0.0000199x(x-2.5069)$$

whence

$$y(3.7608)=0.3989423-0.0000500\cdot 3.7608-$$
$$-0.0000199\cdot 3.7608\cdot(3.7608-2.5069)=0.3986604$$

14.20 INVERSE INTERPOLATION FOR THE CASE OF EQUALLY SPACED POINTS

Suppose we have a function $y=f(x)$ given in tabular form.

The problem of *inverse interpolation* consists in determining a value of the argument x from a given value of the function y.

Let us first examine the case of equally spaced points. Here, the usual method is that of *successive approximations*.

We assume that the function $f(x)$ is monotonic and the given value of y lies between $y_0=f(x_0)$ and $y_1=f(x_1)$.

Replacing the function y by Newton's first interpolation polynomial, we get

$$y=y_0+\frac{\Delta y_0}{1!}q+\frac{\Delta^2 y_0}{2!}q(q-1)+\ \ldots\ +\frac{\Delta^n y_0}{n!}q(q-1)\ \ldots\ (q-n+1)$$

whence $q=\varphi(q)$, where

$$\varphi(q)=\frac{y-y_0}{\Delta y_0}-\frac{\Delta^2 y_0}{2!\,\Delta y_0}q(q-1)-\ \dots\ -\frac{\Delta^n y_0}{n!\,\Delta y_0}q(q-1)\dots(q-n+1)$$

For the initial approximation we take

$$q_0=\frac{y-y_0}{\Delta y_0}$$

Then, applying the method of iteration, we obtain

$$q_m=\varphi(q_{m-1})\qquad (m=1,\ 2,\ \dots) \tag{1}$$

If $f(x)\in C^{(n+1)}[a,\ b]$ where the interval $[a,\ b]$ contains the interpolation points and the spacing h is small, then the process converges:

$$\lim_{m\to\infty} q_m=q$$

where q is the true solution.

Actually, the process of iteration is continued until digits appear which meet the required accuracy; and one assumes $q\approx q_s$, where q_s is the last approximation.

Having found q, we determine x from the formula

$$\frac{x-x_0}{h}=q$$

whence

$$x=x_0+qh$$

Example 1. Using the values of the function $y=\log_{10}x$ given in the table

x	20	25	30
y	1.3010	1.3979	1.4771

find the value of x such that $y=\log_{10}x=1.35$.

Solution. Form the difference table.

TABLE 52

FINITE DIFFERENCES OF THE FUNCTION y

x	y	Δy	$\Delta^2 y$
20	1.3010	969	—177
25	1.3979	792	
30	1.4771		

Assuming $y_0 = 1.3010$, we have

$$q_0 = \frac{y - y_0}{\Delta y_0} = \frac{1.35 - 1.3010}{0.0969} = \frac{490}{969} = 0.506$$

Then, to three decimal places, we successively obtain

$$q_1 = 0.506 - \frac{177}{2 \cdot 969} \cdot 0.506\,(1 - 0.506) = 0.506 - 0.023 = 0.483$$

$$q_2 = 0.506 - \frac{177}{2 \cdot 969} \cdot 0.483\,(1 - 0.483) = 0.506 - 0.023 = 0.483$$

We take

$$q = 0.483$$

whence

$$x = x_0 + qh = 20 + 0.483 \cdot 5 = 22.42$$

By a table of antilogarithms we have $x = 22.39$. The considerable divergence between the computed value and the exact value is due to the large spacing $h = 5$.

We applied the iteration method to solve a problem of inverse interpolation using Newton's first interpolation formula. We could also similarly apply this method to the other interpolation formulas as well: to Newton's second formula, Stirling's formula, Bessel's formula, etc. This is illustrated in the following example.

Example 2. Table 53 contains the values of the probability integral [3]

$$y = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}\, dx$$

For what value of x does the integral y equal $\frac{1}{2}$?

TABLE 53

VALUES OF THE PROBABILITY INTEGRAL

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0.45	0.4754818				
		91737			
0.46	0.4846555		—840		
		90897		—11	
0.47	0.4937452		—851		1
		90046		—10	
0.48	0.5027498		—861		2
		89185		—8	
0.49	0.5116683		—869		
		88316			
0.50	0.5204999				

Solution. We supplement Table 53 with the finite differences of the function y. The closest tabular value of the argument x corresponding to the value of the function $y=\frac{1}{2}$ is $x_0=0.47$. It is convenient here to use Bessel's formula.

We have $x_0=0.47$, $h=0.01$, $y=0.5$.

Substituting these values into (8) of Sec. 14.7 and using the appropriate tabulated values, we obtain

$$0.5=0.4982475+0.0090046p+\frac{p^2-0.25}{2}\left(\frac{-851-861}{2}\right)\cdot 10^{-7}+ \\ +\frac{p(p^2-0\ 25)}{6}(-10)\cdot 10^{-7} \qquad (2)$$

Then, dividing both members of (2) by 0.0090046 and isolating the term containing p to the first power, we get

$$p=0.194623+4.753\cdot 10^{-3}(p^2-0.25)+1.85\cdot 10^{-5}p(p^2-0.25) \qquad (3)$$

For the first approximation of the parameter p we take

$$p^{(1)}=0.194623$$

Putting $p^{(1)}$ in (3), we get the second approximation:

$$p^{(2)}=0.194623+4.753\cdot 10^{-3}[(0.194623)^2-0.25]+ \\ +1.85\cdot 10^{-5}\cdot 0.194623\cdot[(0.194623)^2-0.25]= \\ =0.194623-0.001008-0.000001=0.193614$$

Analogously, putting $p^{(2)}$ in place of p into (3), we get the third approximation:

$$p^{(3)}=0.193612$$

Since the first five decimals coincide, the iteration process is terminated.

Then, we successively find

$$q=p+\frac{1}{2}=0.693612$$

and

$$x=x_0+qh=0.47+0.01\cdot 0.693612=0.47693612$$

This value is correct to the sixth decimal place.

14.21 INVERSE INTERPOLATION FOR THE CASE OF UNEQUALLY SPACED POINTS

The problem of inverse interpolation of a function for the case of unequally spaced values of the argument $x_0, x_1, \ldots, x_n$ can be solved directly by means of Lagrange's interpolation formula. To do this, it suffices to take the variable y as the independent variable and to write a formula expressing x as a function of y (Fig. 64):

$$x=\sum_{i=0}^{n}\frac{(y-y_1)(y-y_2)\dots(y-y_{i-1})(y-y_{i+1})\dots(y-y_n)}{(y_i-y_1)(y_i-y_2)\dots(y_i-y_{i-1})(y_i-y_{i+1})\dots(y_i-y_n)}x_i \tag{1}$$

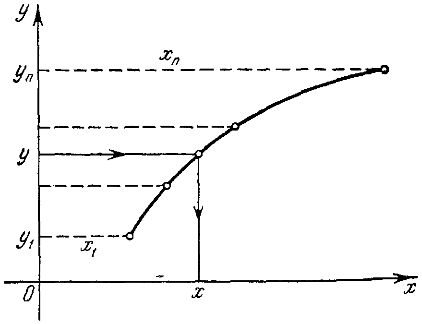


Fig. 64

where $y_i=f(x_i)\,(i=0,\ 1,\ \dots,\ n)$. One could also use Newton's interpolation formula for unequally spaced values of the argument (see Sec. 14.19) taking y as the argument:

$$x=x_0+[y_0,\,y_1](y-y_0)+[y_0,\,y_1,\,y_2](y-y_0)(y-y_1)+\dots$$
$$\dots+[y_0,\,y_1,\,\dots,\,y_n](y-y_0)(y-y_1)\dots(y-y_{n-1}) \tag{2}$$

where $[y_0,\,y_1]$, $[y_0,\,y_1,\,y_2]$, $\dots$, $[y_0,\,y_1,\,\dots,\,y_n]$ are the appropriate divided differences.

Example. Solve Example 2 of Sec. 14.20 with the aid of Lagrange's formula for inverse interpolation.

Solution. We confine ourselves to four values:

$$x_0=0.46,\quad x_1=0.47,\quad x_2=0.48,\quad x_4=0.49$$

Putting

$$u=10^7y-\frac{1}{2}\cdot10^7$$

we have the following table

x	0.46	0.47	0.48	0.49
u	—153 445	—62 548	27 498	116 683

The given value $y=\frac{1}{2}$ corresponds to $u=0$. Using formula (2), where y is replaced by u, we obtain

$$\begin{aligned}x=&\frac{62\,548\cdot(-27\,498)\cdot(-116\,683)}{(-153\,445+62\,548)\cdot(-153\,445-27\,498)\cdot(-153\,445-116\,683)}\cdot 0.46+\\&+\frac{153\,445\cdot(-27\,498)\cdot(-116\,683)}{(-62\,548+153\,445)\cdot(-62\,548-27\,498)\cdot(-62\,548-116\,683)}\cdot 0.47+\\&+\frac{153\,445\cdot 62\,548\cdot(-116\,683)}{(27\,498+153\,445)\cdot(27\,498+62\,548)\cdot(27\,498-116\,683)}\cdot 0.48+\\&+\frac{153\,445\cdot 62\,548\cdot(-27\,498)}{(116\,683+153\,445)\cdot(116\,683+63\,548)\cdot(116\,683-27\,498)}\cdot 0.49=\\&=-0.020779+0.157737+0.369928-0.029950=0.476936\end{aligned}$$

14.22 FINDING THE ROOTS OF AN EQUATION BY INVERSE INTERPOLATION

In conclusion, note that the solution of the equation

$$f(x)=0$$

can be reduced to the problem of inverse interpolation. This requires forming a table of values of the function $y=f(x)$ and constructing a corresponding table of finite differences for the values of x that are close to the root, and then applying the techniques of inverse interpolation to find the value of x that corresponds to $y=0$.

Example. Using the given table of values of the Bessel function $y=J_0(x)$

x	2.4	2.5	2.6
y	0.0025	−0.0484	−0.0968

find the root of the equation $J_0(x)=0$ lying in the interval (2.4, 2.6) to within 10^{-3}.

Solution. Form the difference table (Table 54).

TABLE 54

FINITE DIFFERENCES OF THE BESSEL FUNCTION $y=J_0(x)$

x	y	Δy	$\Delta^2 y$
2.4	0.0025	−509	25
2.5	−0.0484	−484	
2.6	−0.0968		

Putting $y=0$ and $x_0=2.4$, $y_0=0.0025$, we get, on the basis of formula (1) of Sec. 14.20,

$$q_0=\frac{y-y_0}{\Delta y_0}=\frac{0.0025}{0.0509}=0.049,$$

$$q_1=q_0+\frac{\Delta^2 y_0}{2\Delta y_0}q_0(1-q_0)=$$

$$=0.049-\frac{25}{2\cdot 509}\cdot 0.049\cdot 0.951=0.049-0.001=0.048,$$

$$q_2=0.049-\frac{25}{2\cdot 509}\cdot 0.048\cdot 0.952=0.049-0.001=0.048$$

We take

$$q=0.048$$

whence

$$x=x_0+qh=2.4+0.048\cdot 0.1=2.405$$

From the tables,

$$x=2.4048$$

14.23 THE INTERPOLATION METHOD FOR EXPANDING A SECULAR DETERMINANT

Interpolation of functions may be used for expanding a secular (characteristic) determinant (see Chapter 12)

$$D(\lambda)=\det(A-\lambda E)$$

where $A=[a_{ij}]$.

Choose equally spaced points

$$\lambda_0=0,\quad \lambda_1=1,\ \ldots,\ \lambda_n=n$$

and for the determinant $D(\lambda)$ compute the corresponding values

$$D(0)=D_0,\quad D(1)=D_1,\ \ldots,\ D(n)=D_n$$

Forming a horizontal difference table for the sequence of numbers $D(0)$, $D(1)$, $\ldots$, $D(n)$, we find the differences $\Delta^i D(0)$ $(i=0, 1, \ldots, n)$ in the usual manner. From this, using Newton's first interpolation formula, we find the polynomial expression for the secular determinant

$$D(\lambda)=D(0)+\sum_{i=1}^{n}\frac{\Delta^i D(0)}{i!}\lambda(\lambda-1)\ldots(\lambda-i+1) \tag{1}$$

If we put

$$\frac{\lambda(\lambda-1)\ldots(\lambda-i+1)}{i!}=\sum_{m=1}^{i}c_{mi}\lambda^m \quad (i=1,\ 2,\ \ldots) \tag{2}$$

then a few simple manipulations yield *A. A. Markov's formula*:

$$D(\lambda) = D(0) + \sum_{m+1}^{n} \lambda^m \sum_{i=m}^{n} c_{mi} \Delta^i D(0) \tag{3}$$

Tables of coefficients c_{mi} [8] have been compiled to simplify computations by formula (2).

In the more general case, if we take the numbers $\lambda_i = a + ih$ $(i = 0, 1, \ldots, n)$ for the interpolation points, then formula (3) becomes

$$D(\lambda) = D(a) + \sum_{m=1}^{n} (\lambda - a)^m \sum_{i=m}^{n} c_{mi} h^i \Delta^i D(a) \tag{4}$$

Although the *method of interpolation* proposed here requires laborious computations of $n+1$ determinants of order n, it is nevertheless convenient due to its simple computational scheme. What is more, it is applicable to the expansion of a more general type of determinant:

$$F(\lambda) = \det [f_{ij}(\lambda)]$$

where $f_{ij}(\lambda)$ are integral polynomials in λ.

Example. Using the method of interpolation, expand the characteristic determinant

$$D(\lambda) = \begin{vmatrix} 1-\lambda & 2 & 3 & 4 \\ 2 & 1-\lambda & 2 & 3 \\ 3 & 2 & 1-\lambda & 2 \\ 4 & 3 & 2 & 1-\lambda \end{vmatrix}$$

(cf. Sec. 12.3, Example).

Solution. We successively compute $D(i)$ for $i = 0, 1, 2, 3, 4$ to get

$$D(0) = -20, \quad D(1) = -119, \quad D(2) = -308,$$
$$D(3) = -575, \quad D(4) = -884$$

The finite differences $\Delta^i D(0)$ $(i = 0, 1, 2, 3, 4)$ are given in Table 55.

TABLE 55

FINITE DIFFERENCES OF THE NUMBERS $D(\lambda)$

λ	$D(\lambda)$	$\Delta D(\lambda)$	$\Delta^2 D(\lambda)$	$\Delta^3 D(\lambda)$	$\Delta^4 D(\lambda)$
0	$\underline{-20}$	$\underline{-99}$	$\underline{-90}$	$\underline{12}$	$\underline{24}$
1	−119	−189	−78	36	
2	−308	−267	−42		
3	−575	−309			
4	−884				

Since

$$\frac{\lambda}{1!}=\lambda,$$

$$\frac{\lambda(\lambda-1)}{2!}=\frac{\lambda^2}{2}-\frac{\lambda}{2},$$

$$\frac{\lambda(\lambda-1)(\lambda-2)}{3!}=\frac{\lambda^3}{6}-\frac{\lambda^2}{2}+\frac{\lambda}{3},$$

$$\frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)}{4!}=\frac{\lambda^4}{24}-\frac{\lambda^3}{4}+\frac{11\lambda^2}{24}-\frac{\lambda}{4}$$

it follows from formula (2) that

$$c_{11}=1,$$

$$c_{22}=\frac{1}{2},\quad c_{12}=-\frac{1}{2},$$

$$c_{33}=\frac{1}{6},\quad c_{23}=-\frac{1}{2},\quad c_{13}=\frac{1}{3},$$

$$c_{41}=\frac{1}{24},\quad c_{34}=-\frac{1}{4},\quad c_{24}=\frac{11}{24},\quad c_{14}=-\frac{1}{4}$$

whence, using the Markov formula (3), we get

$$\begin{aligned}D(\lambda)&=D(0)+[c_{11}\Delta D(0)+c_{12}\Delta^2D(0)+c_{13}\Delta^3D(0)+c_{14}\Delta^4D(0)]\,\lambda+\\&+[c_{22}\Delta^2D(0)+c_{23}\Delta^3D(0)+c_{24}\Delta^4D(0)]\,\lambda^2+\\&+[c_{33}\Delta^3D(0)+c_{34}\Delta^4D(0)]\,\lambda^2+c_{44}\Delta^4D(0)\,\lambda^4=\\&=-20+\left(-99\cdot1+90\cdot\frac{1}{2}+12\cdot\frac{1}{3}-24\cdot\frac{1}{4}\right)\lambda+\\&+\left(-90\cdot\frac{1}{2}-12\cdot\frac{1}{2}+24\cdot\frac{11}{24}\right)\lambda^2+\left(12\cdot\frac{1}{6}-24\cdot\frac{1}{4}\right)\lambda^3+\\&+24\cdot\frac{1}{24}\lambda^4=-20-56\lambda-40\lambda^2-4\lambda^3+\lambda^4\end{aligned}$$

*14.24 INTERPOLATION OF FUNCTIONS OF TWO VARIABLES

Let the function

$$z=f(x,\ y)$$

be specified on a set of equally spaced points $(x_i,\ y_j)$ $(i,\ j=0,\ 1,\ 2,\ \ldots)$, where

$$x_i=x_0+ih,\quad y_i=y_0+jk$$

and

$$h=\Delta x_i=\text{constant},\quad k=\Delta y_j=\text{constant}$$

For brevity we introduce the notation

$$z_{ij}=f(x_i,\ y_j)$$

The values of the function z may be arranged in a double-entry table (Table 56).

TABLE 56

THE VALUES OF A FUNCTION OF TWO VARIABLES

x y	x_0	x_1	x_2	...
y_0	z_{00}	z_{10}	z_{20}	...
y_1	z_{01}	z_{11}	z_{21}	...
y_2	z_{02}	z_{12}	z_{22}	...
...	...	...	...	...

Interpolation of a function of two variables

$$z = f(x, y)$$

that is, the finding of nontabulated values, may be performed successively with respect to each variable x and y separately. For example, let it be required to find the value

$$\bar{z} = f(\bar{x}, \bar{y})$$

Interpolating in appropriate fashion the chosen functions of one variable x,

$$f_k(x) = f(x, y_k)$$

where $y_k \approx \bar{y}$, we find the values $f_k(\bar{x})$. For this purpose one uses the corresponding rows of the double-entry table. Regarding the values $f_k(\bar{x}) = f(\bar{x}, y_k)$ thus obtained as the values of the function $f(\bar{x}, y)$ of one variable y, we find the desired value $f(\bar{x}, \bar{y}) = \bar{z}$ by means of one of the interpolation formulas.

The interpolation can also be carried out in the reverse order.

Example. The values of the "aftereffect" function

$$f(x, y) = \int_{-\infty}^{+\infty} e^{-y^2z^2 - z - xe^{-z}}\, dz$$

are given in the following table (see Jahnke and Emde's *Funk-*

tionentafeln):

$y \backslash x$	0.4	0.7	1.0
0.00	2.500	1.429	1.000
0.05	2.487	1.419	0.995
0.10	2.456	1.400	0.981

Find $f(0.5,\ 0.03)$.

Solution. Form the tables 57a, 57b and 57c using the rows of the given double-entry table.

$y=0$

TABLE 57a

x	f	Δf	$\Delta^2 f$
0.4	2.500	—1.071	0.642
0.7	1.429	—0.429	
1.0	1.000		

$y=0.05$

TABLE 57b

x	f	Δf	$\Delta^2 f$
0.4	2.487	—1.068	0.644
0.7	1.419	—0.424	
1.0	0.995		

$y=0.10$

TABLE 57c

x	f	Δf	$\Delta^2 f$
0.4	2.456	—1.056	0.637
0.7	1.400	—0.419	
1.0	0.981		

Since for these tables

$$h=0.7-0.4=0.3$$

then, putting $x_0=0.4$, we have,

$$q=\frac{x-x_0}{h}=\frac{0.5-0.4}{0.3}=\frac{1}{3}$$

whence, using Newton's first interpolation formula, we successively obtain

$$f_0=f(0.5,\ 0)=2.500-\frac{1}{3}\cdot 1.071+\frac{\frac{1}{3}\left(-\frac{2}{3}\right)}{2}\cdot 0.642=2.072,$$

$$f_1=f(0.5,\ 0.05)=2.487-\frac{1}{3}\cdot 1.068-\frac{1}{9}\cdot 0.644=2.069,$$

$$f_2=f(0.5,\ 0.10)=2.456-\frac{1}{3}\cdot 1.056-\frac{1}{9}\cdot 0.637=2.033$$

We now form a table of the values thus found (Table 58).

TABLE 58

y	f	Δf	$\Delta^2 f$
0	2.072	—0.003	—0.033
0.05	2.069	—0.036	
0.10	2.033		

Assuming $k=0.05-0=0.05$ and $y_0=0$, we get

$$q'=\frac{0.03-0}{0.05}\doteq\frac{3}{5}$$

whence

$$f(0.5,\ 0.03)=2.072-\frac{3}{5}\cdot 0.003+\frac{\frac{3}{5}\cdot\left(-\frac{2}{5}\right)}{2}\cdot(-0.033)=2.074$$

*14.25 DOUBLE DIFFERENCES OF HIGHER ORDER

For a function $z=f(x,\ y)$ given by a double-entry table $\{z_{ij}\}$ we can define the *partial finite differences*

$$\Delta_x z_{ij}=z_{i+1,\ j}-z_{ij} \quad \text{and} \quad \Delta_y z_{ij}=z_{i,\ j+1}-z_{ij}$$

Repeating these operations, we obtain *double differences of higher orders*:

$$\Delta^{m+n}z_{ij}=\Delta^{m+n}_{x^m y^n}z_{ij}=\Delta^m_{x^m}(\Delta^n_{y^n}z_{ij})=\Delta^n_{y^n}(\Delta^m_{x^m}z_{ij}),$$

where we have set $\Delta^{0+0}z_{ij}=z_{ij}$. For instance

$$\Delta^{1+2}z_{ij}=\Delta_x(\Delta^2_{yy}z_{ij})=\Delta_x(z_{i,\ j+2}-2z_{i,\ j+1}+z_{ij})=$$

$$=(z_{i+1,\ j+2}-2z_{i+1,\ j+1}+z_{i+1,\ j})-(z_{i,\ j+2}-2z_{i,\ j+1}+z_{ij})$$

*14.26 NEWTON'S INTERPOLATION FORMULA FOR A FUNCTION OF TWO VARIABLES

Using the differences of a function of two variables $z=f(x, y)$, we can construct an interpolation polynomial similar to Newton's interpolation polynomial. Let $P(x, y)$ be an integral polynomial such that

$$\Delta_{x^m y^n}^{m+n} P(x_0, y_0) = \Delta^{m+n} z_{00} \tag{1}$$

$(m, n=0, 1, 2, \ldots)$. Suppose that $P(x, y)$ is expanded in terms of the generalized powers of the differences $x-x_0$ and $y-y_0$; that is,

$$\begin{aligned} P(x, y) = c_{00} + c_{10}(x-x_0) + c_{01}(y-y_0) + c_{20}(x-x_0)(x-x_1) + \\ + c_{11}(x-x_0)(y-y_0) + c_{02}(y-y_0)(y-y_1) + \ldots \end{aligned} \tag{2}$$

Putting $x=x_0$ and $y=y_0$, we have, by virtue of (1),

$$P(x_0, y_0) = z_{00} = c_{00}$$

We form first differences for $P(x, y)$ to get

$$\Delta_x P(x, y) = c_{10}h + 2c_{20}h(x-x_0) + c_{11}h(y-y_0) + \ldots$$

and

$$\Delta_y P(x, y) = c_{01}k + c_{11}k(x-x_0) + 2c_{02}k(y-y_0) + \ldots$$

whence, putting $x=x_0$ and $y=y_0$, we have, on the basis of (1),

$$\Delta_x P(x_0, y_0) = \Delta^{1+0} z_{00} = c_{10}h$$

and

$$\Delta_y P(x_0, y_0) = \Delta^{0+1} z_{00} = c_{01}k$$

That is,

$$c_{10} = \frac{\Delta^{1+0} z_{00}}{h}, \qquad c_{10} = \frac{\Delta^{0+1} z_{00}}{k}$$

Then, computing the finite differences of second order for the polynomial $P(x, y)$, we find

$$\begin{aligned} \Delta_{xx} P(x, y) &= 2!\, c_{20}h^2 + \ldots, \\ \Delta_{xy} P(x, y) &= c_{11}hk + \ldots, \\ \Delta_{yy} P(x, y) &= 2!\, c_{02}k^2 + \ldots \end{aligned}$$

From this we obtain, for $x=x_0$ and $y=y_0$,

$$\begin{aligned} \Delta_{xx} P(x_0, y_0) &= \Delta^{2+0} z_{00} = 2!\, c_{20}h^2, \\ \Delta_{xy} P(x_0, y_0) &= \Delta^{1+1} z_{00} = c_{11}hk, \\ \Delta_{yy} P(x_0, y_0) &= \Delta^{0+2} z_{00} = 2!\, c_{02}k^2 \end{aligned}$$

Thus

$$c_{20}=\frac{1}{2!}\cdot\frac{\Delta^{2+0}z_{00}}{h^2},\qquad c_{11}=\frac{\Delta^{1+1}z_{00}}{hk},\qquad c_{02}=\frac{1}{2!}\cdot\frac{\Delta^{0+2}z_{00}}{k^2}$$

The subsequent coefficients of the expansion (2) are found in a similar manner. Substituting the values of the coefficients thus found into formula (2), we get an *interpolation polynomial for a function of two variables*:

$$\begin{aligned}P(x,\ y)=z_{00}&+\left[\frac{\Delta^{1+0}z_{00}}{h}(x-x_0)+\frac{\Delta^{0+1}z_{00}}{k}(y-y_0)\right]+\\&+\frac{1}{2!}\left[\frac{\Delta^{2+0}z_{00}}{h^2}(x-x_0)^{[2]}+2\cdot\frac{\Delta^{1+1}z_{00}}{hk}(x-x_0)(y-y_0)+\right.\\&\left.+\frac{\Delta^{0+2}z_{00}}{k^2}(y-y_0)^{[2]}\right]+\dots\end{aligned}\tag{3}$$

When interpolating the function $f(x,\ y)$ we assume

$$f(x,\ y)\approx P(x,\ y)$$

To simplify computations, it is common to introduce the variables

$$\frac{x-x_0}{h}=p,\qquad \frac{y-y_0}{k}=q$$

Then

$$\frac{x-x_1}{h}=p-1,\qquad \frac{y-y_1}{k}=q-1$$

and so on. Then formula (3) becomes

$$\begin{aligned}z\approx z_{00}+(p\Delta^{1+0}z_{00}+q\Delta^{0+1}z_{00})+\frac{1}{2!}[p(p-1)\,\Delta^{2+0}z_{00}+\\+2pq\Delta^{1+1}z_{00}+q(q-1)\,\Delta^{0+2}z_{00}]+\dots\end{aligned}\tag{4}$$

where

$$x=x_0+ph,\qquad y=y_0+qk$$

If one puts $p=0$ or $q=0$, then (4) becomes the corresponding interpolation formula of Newton.

Example. Using the interpolation formula (4), find $f=f(0.5,\ 0.03)$ for the function $f(x,\ y)$ considered in the example of Sec. 14.24.

Solution. Taking $x_0=0.4$, $y_0=0$, form tables of first differences (Tables 59a and 59b) for f.

TABLE 59a

	$\Delta^{1+0}f_{0j}$	$\Delta^{1+0}f_{1j}$
$j=0$	—1.071	—0.429
$j=1$	—1.068	—0.424
$j=2$	—1.056	—0.419

TABLE 59b

	$i=0$	$i=1$	$i=2$
$\Delta^{0+1}f_{i0}$	—0.013	—0.010	—0.005
$\Delta^{0+1}f_{i1}$	—0.031	—0.019	—0.014

From this we find the second differences

$$\Delta^{2+0}f_{00}=\Delta^{1+0}f_{10}-\Delta^{1+0}f_{00}=-0.429-(-1.071)=0.642,$$
$$\Delta^{1+1}f_{00}=\Delta^{1+0}f_{01}-\Delta^{1+0}f_{00}=-1.068-(-1.071)=0.003$$

or

$$\Delta^{1+1}f_{00}=\Delta^{0+1}f_{10}-\Delta^{0+1}f_{00}=-0.010-(-0.013)=0.003,$$
$$\Delta^{0+2}f_{00}=\Delta^{0+1}f_{01}-\Delta^{0+1}f_{00}=-0.031-(-0.013)=-0.018$$

Since

$$p=\frac{x-x_0}{h}=\frac{1}{3}, \qquad q=\frac{y-y_0}{k}=\frac{3}{5}$$

then, using formula (4), we get

$$f=2.500+\frac{1}{3}\cdot(-1.071)+\frac{3}{5}\cdot(-0.013)+\frac{1}{2}\left[\frac{1}{3}\cdot\left(-\frac{2}{3}\right)\cdot 0.642+\right.$$
$$\left.+2\cdot\frac{1}{3}\cdot\frac{3}{5}\cdot 0.003+\frac{3}{5}\cdot\left(-\frac{2}{5}\right)(-0.018)\right]=$$
$$=2.500-0.357-0.0078-0.0713+0.0006+0.0021=2.067$$

Comparing this with the answer $f=2.074$ obtained by the first method, we see that the thousandths digits are unreliable.

REFERENCES FOR CHAPTER 14

[1] *E. T. Whittaker and G. Robinson, The Calculus of Observations*, 1944, Chapter I.

[2] *V. L. Goncharov, The Theory of Interpolation and Approximation of Functions*, 1934, Chapter I, Secs. 18-21 (in Russian).

[3] *James B. Scarborough, Numerical Mathematical Analysis*, 1955, Chapter IV, Part II.

[4] *V. M. Bradis, The Theory and Practice of Computations*, 1935, Chapter IX (in Russian).

[5] *E. Milne, Numerical Calculus*, 1949, Chapters III and VI.

[6] *E. Ya. Remez, General Computational Methods of Chebyshev Approximation*, 1957, Part I, Chapter I (in Russian).

[7] *N. A. Lednev (editor), Mathematical Practice Session Devoted to Computing Instruments and Machines*, 1959, Chapter III (in Russian).

[8] *V. N. Faddeyeva, Computational Methods of Linear Algebra*, 1950, Chapter III, Sec. 27 (in Russian).

Chapter 15

APPROXIMATE DIFFERENTIATION

15.1 STATEMENT OF THE PROBLEM

In the solution of practical problems, one often has to find derivatives of indicated orders of a function $y=f(x)$ given in tabular form. It is also possible that due to the complexity of the analytic expression of the function $f(x)$, direct differentiation would be involved. In such cases, one usually resorts to *approximate differentiation.*

To derive formulas for approximate differentiation we replace the given function $f(x)$, on the interval $[a, b]$ that interests us, by an interpolating function $P(x)$ (mostly by a polynomial) and then set

$$f'(x)=P'(x) \tag{1}$$

for

$$a \leqslant x \leqslant b$$

We do the same when seeking higher-order derivatives of the function $f(x)$.

If we know the error

$$R(x)=f(x)-P(x)$$

in the interpolating function $P(x)$, then the error of the derivative $P'(x)$ is given by the formula

$$r(x)=f'(x)-P'(x)=R'(x) \tag{2}$$

which means that *the error of the derivative of an interpolating function is equal to the derivative of the error in that function.* The same holds true for higher-order derivatives as well.

It is well to note that, generally speaking, approximate differentiation is a less exact operation than interpolation. Indeed, the closeness of the ordinates of two curves

$$y=f(x) \quad \text{and} \quad Y=P(x)$$

on the interval $[a, b]$ does not yet guarantee closeness of their derivatives $f'(x)$ and $P'(x)$ on that interval; that is, it does not guarantee a small divergence of the slopes of the tangents to the curves at hand, given the same values of the argument (Fig. 65).

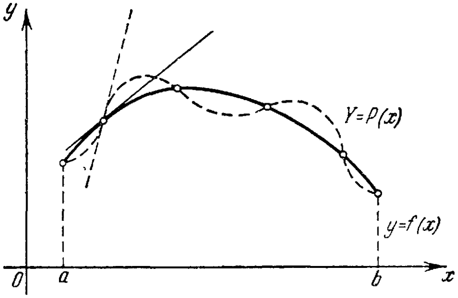


Fig. 65

15.2 FORMULAS OF APPROXIMATE DIFFERENTIATION BASED ON NEWTON'S FIRST INTERPOLATION FORMULA

Suppose we have a function $y(x)$ specified at equally spaced points x_i $(i=0, 1, 2, \ldots, n)$ of an interval $[a, b]$ by means of the values $y_i=f(x_i)$. In order to find on $[a, b]$ the derivatives $y'=f'(x)$, $y''=f''(x)$, etc.,[1] we replace the function y by Newton's interpolation polynomial constructed for a set of points x_0, $x_1, \ldots, x_k$ $(k\leqslant n)$.

We have

$$y(x)=y_0+q\Delta y_0+\frac{q(q-1)}{2!}\Delta^2 y_0+\frac{q(q-1)(q-2)}{3!}\Delta^3 y_0+ \\ +\frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4 y_0+\ldots \quad (1)$$

where

$$q=\frac{x-x_0}{h} \quad \text{and} \quad h=x_{i+1}-x_i \quad (i=0, 1, \ldots)$$

Multiplying the binomials together, we get

$$y(x)=y_0+q\Delta y_0+\frac{q^2-q}{2}\Delta^2 y_0+\frac{q^3-3q^2+2q}{6}\Delta^3 y_0+ \\ +\frac{q^4-6q^3+11q^2-6q}{24}\Delta^4 y_0+\ldots \quad (1')$$

Since

$$\frac{dy}{dx}=\frac{dy}{dq}\cdot\frac{dq}{dx}=\frac{1}{h}\frac{dy}{dq}$$

---

[1] Quite naturally, it must be known beforehand that the appropriate derivatives of $f(x)$ exist, otherwise the computations will be illusory.

it follows that

$$y'(x)=\frac{1}{h}\left[\Delta y_0+\frac{2q-1}{2}\Delta^2 y_0+\frac{3q^2-6q+2}{6}\Delta^3 y_0+\right.$$
$$\left.+\frac{2q^3-9q^2+11q-3}{12}\Delta^4 y_0+\ldots\right] \quad (2)$$

Similarly, since

$$y''(x)=\frac{d(y')}{dx}=\frac{d(y')}{dq}\cdot\frac{dq}{dx}$$

it follows that

$$y''(x)=\frac{1}{h^2}\left[\Delta^2 y_0+(q-1)\Delta^3 y_0+\frac{6q^2-18q+11}{12}\Delta^4 y_0+\ldots\right] \quad (3)$$

If necessary, the same method may be used to compute the derivatives of any order of the function $y(x)$.

Note that when seeking the derivatives $y'(x)$, $y''(x)$, ... at a fixed point x, one should choose for x_0 the closest tabular value of the argument.

It is sometimes required to find the derivatives of y at basic tabulated points x_i. In this case the formulas of numerical differentiation are simplified. Since each tabular value may be taken as the initial value, we put $x=x_0$, $q=0$; then we have

$$y'(x_0)=\frac{1}{h}\left(\Delta y_0-\frac{\Delta^2 y_0}{2}+\frac{\Delta^3 y_0}{3}-\frac{\Delta^4 y_0}{4}+\frac{\Delta^5 y_0}{5}-\ldots\right) \quad (4)$$

and

$$y''(x_0)=\frac{1}{h^2}\left(\Delta^2 y_0-\Delta^3 y_0+\frac{11}{12}\Delta^4 y_0-\frac{5}{6}\Delta^5 y_0+\ldots\right) \quad (5)$$

If $P_k(x)$ is Newton's interpolation polynomial containing the differences Δy_0, $\Delta^2 y_0$, ..., $\Delta^k y_0$ and

$$R_k(x)=y(x)-P_k(x)$$

is the corresponding error, then the error in determining the derivative is

$$R'_k(x)=y'(x)-P'_k(x)$$

As we know (Sec. 14.15),

$$R_k(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_k)}{(k+1)!}y^{(k+1)}(\xi)=$$
$$=h^{k+1}\frac{q(q-1)\ldots(q-k)}{(k+1)!}y^{(k+1)}(\xi)$$

where ξ is an intermediate number between the values x_0, x_1, ...

..., x_k, x. For this reason, assuming that $y(x) \in C^{(k+2)}$, we get

$$R'_k(x) = \frac{dR_k}{dq} \cdot \frac{dq}{dx} = \frac{h^k}{(k+1)!} \left\{ y^{(k+1)}(\xi) \frac{d}{dq}[q(q-1)\dots(q-k)] + \right.$$
$$\left. + q(q-1)\dots(q-k) \frac{d}{dq}[y^{(k+1)}(\xi)] \right\}$$

Furthermore, if we suppose $\frac{d}{dq}[y^{(k+1)}(\xi)]$ to be bounded and take into account that $\frac{d}{dq}[q(q-1)\dots(q-k)]_{q=0} = (-1)^k k!$, then for $x = x_0$ and, hence, for $q = 0$, we will have

$$R'_k(x_0) = (-1)^k \frac{h^k}{k+1} y^{(k+1)}(\xi) \tag{6}$$

Since in many cases it is difficult to estimate $y^{(k+1)}(\xi)$, for small h we set

$$y^{(k+1)}(\xi) \approx \frac{\Delta^{k+1} y_0}{h^{k+1}}$$

and, hence,

$$R'_k(x_0) \approx \frac{(-1)^k}{h} \frac{\Delta^{k+1} y_0}{k+1} \tag{7}$$

In similar fashion we can find the error $R''_k(x_0)$ for the second derivative $y''(x_0)$.

Example 1. Find $y'(50)$ of the tabulated function $y = \log_{10} x$ (Table 60).

TABLE 60

VALUES OF THE FUNCTION $y = \log_{10} x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
50	1.6990	414	—36	5
55	1.7404	378	—31	
60	1.7782	347		
65	1.8129			

Solution. Here $h = 5$. We supplement Table 60 with columns of finite differences (as usual, decimal orders are not indicated; they are determined by the decimal orders of the values of the function).

Using the first row of the table, on the basis of formula (4), we get, to within third differences,

$$y'(50) = \frac{1}{5}(0.0414 + 0.0018 + 0.0002) = 0.0087$$

To estimate the accuracy of this value, note that since the above-tabulated function is $y=\log_{10}x$,

$$y'_x=\frac{M}{x}=\frac{0.43429}{x}$$

Hence

$$y'(50)=\frac{0.43429}{50}=0.0087$$

Thus, the results coincide to the fourth decimal place.

Example 2. The path $y=f(t)$ traversed in time t by a point moving in a straight line is given by the following table [1]:

i	Time t_i in sec	Path $y(t_i)$ in cm
0	0.00	0.000
1	0.01	1.519
2	0.02	6.031
3	0.03	13.397
4	0.04	23.396
5	0.05	35.721
6	0.06	50.000
7	0.07	65.798
8	0.08	82.635
9	0.09	100.000

Using finite differences up to order five inclusive, we get the approximate velocity $V=\frac{dy}{dt}$ and the acceleration $W=\frac{d^2y}{dt^2}$ of the point for times $t=0$, 0.01, 0.02, 0.03, 0.04.

Solution. Form the difference table (Table 61).

Setting $h=0.01$ and applying formulas (4) and (5), we get approximate values of the velocity V (cm/sec) and acceleration W (cm/sec^2). For example,

$$V(0)=100(1.519-1.496-0.046+0.020-0.001)=-0.4 \text{ cm/sec}$$

$$W(0)=10.000(2.993+0.139-0.075+0.003)=30{,}600 \text{ cm/sec}^2$$

The values of V and W are given in Table 62.

TABLE 61

FINITE DIFFERENCES OF THE FUNCTION $y = f(t)$

i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$
0	1.519	2.993	—0.139	—0.082	—0.004
1	4.512	2.854	—0.221	—0.086	0.021
2	7.366	2.633	—0.307	—0.065	0.002
3	9.999	2.326	—0.372	—0.063	0.018
4	12.325	1.954	—0.435	—0.045	0.014
5	14.279	1.519	—0.480	—0.031	—
6	15.798	1.039	—0.511	—	
7	16.837	0.528	—		
8	17.365	—			
9	—				

TABLE 62

VALUES OF VELOCITY V AND ACCELERATION W FOR THE LAW OF MOTION $y = f(t)$

t	V	W	$\tilde{V}$	$\tilde{W}$
0.00	0.4	30,600	0.00	30,462
0.01	303.6	29,780	303.08	30,001
0.02	596.3	28,780	596.98	28,625
0.03	873.2	26,250	872.66	26,381
0.04	1121.7	23,360	1121.9	23,340

It will be noted that the tabulated law of motion is given by the formula

$$y = 100\left(1 - \cos\frac{50\pi t}{9}\right)$$

whence

$$V = \frac{dy}{dt} = \frac{5000\pi}{9}\sin\frac{50\pi t}{9}$$

and

$$W = \frac{d^2y}{dt^2} = \frac{250\,000\pi^2}{81}\cos\frac{50\pi t}{9}$$

By way of comparison, the exact values, $\tilde{V}$ and $\tilde{W}$, are given on the right-hand side of Table 62.

It may be noted that it is also possible to derive formulas for approximate differentiation by proceeding from Newton's second interpolation formula.

15.3 FORMULAS OF APPROXIMATE DIFFERENTIATION BASED ON STIRLING'S FORMULA

The formulas of numerical differentiation derived in Sec. 15.2 for a function y at a point $x=x_0$ have the disadvantage that they only employ one-sided values of the function for $x>x_0$. A relatively higher accuracy is assured by symmetric formulas of differentiation which take into account the values of the given function y both for $x>x_0$ and for $x<x_0$. These formulas are ordinarily called *central formulas (formulas for central derivatives)*. We confine ourselves to the derivation of one of them and take Stirling's interpolation formula as a basis.

Let $\ldots, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \ldots$ be a set of equally spaced points with $x_{i+1}-x_i=h$ the spacing and $y_i=f(x_i)$ the corresponding values of the given function $y=f(x)$. Setting

$$q=\frac{x-x_0}{h}$$

and replacing the function y approximately by the Stirling interpolation polynomial, we get

$$y(x)=y_0+q\Delta y_{-\frac{1}{2}}+\frac{q^2}{2!}\Delta^2 y_{-1}+\frac{q(q^2-1)}{3!}\Delta^3 y_{-\frac{3}{2}}+$$
$$+\frac{q^2(q^2-1)}{4!}\Delta^4 y_{-2}+\frac{q(q^2-1)(q^2-2^2)}{5!}\Delta^5 y_{-\frac{5}{2}}+$$
$$+\frac{q^2(q^2-1)(q^2-2^2)}{6!}\Delta^6 y_{-3}+\ldots, \qquad (1)$$

where for brevity we introduce the notation

$$\Delta y_{-\frac{1}{2}}=\frac{\Delta y_{-1}+\Delta y_0}{2},$$

$$\Delta^3 y_{-\frac{3}{2}}=\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2},$$

$$\Delta^5 y_{-\frac{5}{2}}=\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2}$$

and so forth.

From formula (1), noting that

$$\frac{dq}{dx}=\frac{1}{h},$$

we get

$$y'(x)=\frac{1}{h}\left(\Delta y_{-\frac{1}{2}}+q\Delta^2 y_{-1}+\frac{3q^2-1}{6}\Delta^3 y_{-\frac{3}{2}}+\frac{2q^3-g}{12}\Delta^4 y_{-2}+\right.$$
$$\left.+\frac{5q^4-15q^2+4}{120}\Delta^5 y_{-\frac{5}{2}}+\frac{3q^5-10q^3+4q}{360}\Delta^6 y_{-3}+\dots\right), \qquad (2)$$

$$y''(x)=\frac{1}{h^2}\left(\Delta^2 y_{-1}+q\Delta^3 y_{-\frac{3}{2}}+\frac{6q^2-1}{12}\Delta^4 y_{-2}+\right.$$
$$\left.+\frac{2q^3-3q}{12}\Delta^4 y_{-\frac{5}{2}}+\frac{15q^4-30q^2+4}{360}\Delta^6 y_{-3}+\dots\right) \qquad (2')$$

In particular, setting $q=0$, we have

$$y'(x_0)=\frac{1}{h}\left(\Delta y_{-\frac{1}{2}}-\frac{1}{6}\Delta^3 y_{-\frac{3}{2}}+\frac{1}{30}\Delta^5 y_{-\frac{5}{2}}+\dots\right) \qquad (3)$$

and

$$y''(x_0)=\frac{1}{h^2}\left(\Delta^2 y_{-1}-\frac{1}{12}\Delta^4 y_{-2}+\frac{1}{90}\Delta^6 y_{-3}+\dots\right) \qquad (3')$$

Example 1. Find $y'(1)$ and $y''(1)$ for the function $y=y(x)$ given by Table 63.

TABLE 63

VALUES OF THE FUNCTION $y=y(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0.96	0.7825361				
		—86029			
0.98	0.7739332		—1326		
		$\underline{-87355}$		$\underline{25}$	
1.00	0.7651977		$\underline{\underline{-1301}}$		$\underline{\underline{1}}$
		$\underline{-88656}$		$\underline{26}$	
1.02	0.7563321		—1275		
		—89931			
1.04	0.7473390				

Solution. Forming the differences for the function y (Table 63) and utilizing the underlined terms, we get, on the basis of formula (3),

$$y'(1)=\frac{1}{0.02}\left(-\frac{87355+88656}{2}\cdot 10^{-7}-\right.$$
$$\left.-\frac{1}{6}\cdot\frac{25+26}{2}\cdot 10^{-7}+\frac{1}{30}\cdot 1\cdot 10^{-7}\right)=$$
$$=-50\cdot(88005.5+4.2+0)\cdot 10^{-7}=-0.4400485$$

As a check, note that the tabulated function is Bessel's function of order zero $y=J_0(x)$.

As we know,

$$J_0'(1)=-J_1(x)\big|_{x=1}=-0.4400506$$

Similarly, using the twice underlined terms and applying formula (3′), we have

$$y''(1)=\frac{1}{0.02^2}\cdot\left(-1301\cdot 10^{-7}-\frac{1}{12}\cdot 1\cdot 10^{-7}\right)=$$
$$=-2500\cdot 1301\cdot 10^{-7}=-3.2525\cdot 10^{-1}=-0.325250$$

By way of comparison, we give the exact value obtained on the basis of relations between the Bessel functions:

$$y''(1)=J_0''(1)=J_1(1)-J_0(1)=0.4400506-0.7651977=-0.325147$$

Thus, finding the second derivative numerically is, generally, a less reliable operation than finding the first derivative.

Note. It is sometimes required to find the extremum of a differentiable function $y=y(x)$ given in tabular form. It is then necessary that the equation $y'(\tilde{x})=0$ hold at the point of the extremum $\tilde{x}$. Equating the derivative $y'(x)$ to zero in formula (2), we find the appropriate value of q by the method of successive approximations. From this,

$$\tilde{x}=x_0+qh$$

and the value of $\tilde{y}$ is computed from formula (1) or from some other interpolation formula. The value of $\tilde{y}$ thus found is the extremum of the function if the second difference $\Delta^2 y$ preserves constant sign in the neighbourhood of the point $\tilde{x}$.

Example 2. Find the zero of the derivative of the function $y=J_1(x)$ given by Table 64.

We supplement Table 64 with finite differences of the function y.

Solution. We take $x_0=1.84$. Using the underlined differences, we get, on the basis of formula (2),

$$0=\frac{918-723}{2}+q(-1641)+\frac{3q^2-1}{6}\cdot\frac{2+4}{2}$$

or

$$0=97-1641q+\frac{3}{2}q^2$$

From this,

$$q=\frac{97}{1641}+\frac{1}{1094}q^2 \qquad (4)$$

TABLE 64

THE VALUES OF THE FUNCTION $y = J_1(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1.80	0.5815170			
		2561		
1.82	0.5817731		−1643	
		918		2
1.84	0.5818649		−1641	
		−723		4
1.86	0.5817926		−1637	
		−2360		2
1.88	0,5815566		−1635	
		−3995		
1.90	0.5811571			

Dropping the small nonlinear term, we get the first approximation:

$$q^{(1)} = \frac{97}{1641} = 5.911 \cdot 10^{-2}$$

Improving this value, we obtain the second approximation from formula (4):

$$q^{(2)} = q^{(1)} + \frac{1}{1094}\,[q^{(1)}]^2 = 5.911 \cdot 10^{-2} + \frac{1}{1094} \cdot 3.494 \cdot 10^{-3} =$$
$$= 5.911 \cdot 10^{-2} + 3.2 \cdot 10^{-6} = 5.911 \cdot 10^{-2}$$

Hence we can put

$$q = 0.05911$$

whence

$$x = x_0 + qh = 1.84 + 0.05911 \cdot 0.02 = 1.8411822$$

Thus

$$J_1'(1.8411822) = 0$$

15.4 FORMULAS OF NUMERICAL DIFFERENTIATION FOR EQUALLY SPACED POINTS

Let the points $x_0, x_1, x_2, \ldots, x_n$ be equally spaced; that is,

$$x_{i+1} - x_i = h \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n-1)$$

and for the function $y = y(x)$ let the values $y_i = y(x_i)$ $(i = 0, 1, \ldots, n)$ be known. For this set of points x_i construct Lagrange's interpolation polynomial (Sec. 14.12):

$$L_n(x)=\sum_{i=0}^{n}\frac{\Pi_{n+1}(x)\,y_i}{(x-x_i)\,\Pi'_{n+1}(x_i)}$$

where

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ \ldots\ (x-x_n)$$

Then

$$L_n(x_i)=y_i \qquad (i=0,\ 1,\ \ldots,\ n)$$

Putting

$$\frac{x-x_0}{h}=q$$

we get

$$\Pi_{n+1}(x)=h^{n+1}q(q-1)\ldots(q-n)=h^{n+1}q^{[n+1]}$$

and

$$\begin{aligned}\Pi'_{n+1}(x_i)&=(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)=\\&=h^n i(i-1)\ldots1(-1)\ldots[-(n-i)]=\\&=(-1)^{n-i}h^n_i!\ (n-i)! \qquad (1)\end{aligned}$$

Hence, for the Lagrange polynomial $L_n(x)$ we have the expression

$$L_n(x)=\sum_{i=0}^{n}\frac{(-1)^{n-i}y_i}{i!\,(n-i)!}\cdot\frac{q^{[n+1]}}{q-i} \qquad (2)$$

From this, noting that

$$\frac{dx}{dq}=h$$

we get

$$y'(x)\approx L'_n(x)=\frac{1}{h}\sum_{i=0}^{n}\frac{(-1)^{n-i}y_i}{i!\,(n-i)!}\frac{d}{dq}\left\{\frac{q^{[n+1]}}{q-i}\right\} \qquad (3)$$

The higher order derivatives of the given function $y(x)$ may be found in similar fashion. To estimate the error

$$r_n(x)=y'(x)-L'_n(x)$$

we take advantage of the familiar formula for the error of interpolation formula (2) (Sec. 14.14):

$$R_n(x)=y(x)-L_n(x)=\frac{y^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x) \qquad (4)$$

where $\xi=\xi(x)$ is an intermediate value between the points x_0, x_1, $\ldots$, x_n and x.

Assuming that $y(x)\in C^{(n+2)}$, we derive

$$\begin{aligned}r_n(x)&=R'_n(x)=\\&=\frac{1}{(n+1)!}\left\{y^{(n+1)}(\xi)\,\Pi'_{n+1}(x)+\Pi_{n+1}(x)\frac{d}{dx}\left[y^{(n+1)}(\xi)\right]\right\}\end{aligned}$$

From this, taking into account formula (1) and assuming $\frac{d}{dx}[y^{(n+1)}(\xi)]$ to be bounded, we get the error of the derivative at the points:

$$R_n'(x_i) = (-1)^{n-i} h^n \frac{i!\,(n-i)!}{(n+1)!} y^{(n+1)}(\xi) \tag{5}$$

where ξ is a value lying between x_0 and x_n.

I. Let us perform the computations for $n=2$ (three points). From formula (2) we get

$$L_2(x) = \frac{1}{2} y_0 (q-1)(q-2) - y_1 q (q-2) + \frac{1}{2} y_2 q (q-1)$$

whence, noting that $\frac{dx}{dq} = h$, we have

$$y'(x) \approx L_2'(x) = \frac{1}{h}\left[\frac{1}{2} y_0 (2q-3) - y_1 (2q-2) + \frac{1}{2} y_2 (2q-1)\right]$$

In particular, for the derivatives

$$y'(x_i) = y'_i \, (i = 0,\ 1,\ 2)$$

we obtain the following expressions:

$$y_0' = \frac{1}{2h}(-3y_0 + 4y_1 - y_2),$$

$$y_1' = \frac{1}{2h}(-y_0 + y_2),$$

$$y_2' = \frac{1}{2h}(y_0 - 4y_1 + 3y_2)$$

with corresponding errors:

$$r_0 = \frac{1}{3} h^2 y'''(\xi_0),$$

$$r_1 = -\frac{1}{6} h^2 y'''(\xi_1),$$

$$r_2 = \frac{1}{3} h^2 y'''(\xi_2)$$

We give, without proof, the differentiation formulas for four and five points [3], the validity of which the reader can easily check by himself.

II. $n=3$ (four points):

$$y_0' = \frac{1}{6h}(-11y_0 + 18y_1 - 9y_2 + 2y_3) - \frac{h^3}{4} y^{(4)}(\xi),$$

$$y_1' = \frac{1}{6h}(-2y_0 - 3y_1 + 6y_2 - y_3) + \frac{h^3}{12} y^{(4)}(\xi),$$

$$y'_2 = \frac{1}{6h}(y_0 - 6y_1 + 3y_2 + 2y_3) - \frac{h^3}{12} y^{(4)}(\xi),$$

$$y'_3 = \frac{1}{6h}(-2y_0 + 9y_1 - 18y_2 + 11y_3) + \frac{h^3}{4} y^{(4)}(\xi)$$

III. $n = 4$ (five points):

$$y'_0 = \frac{1}{12h}(-25y_0 + 48y_1 - 36y_2 + 16y_3 - 3y_4) + \frac{h^4}{5} y^{(5)}(\xi),$$

$$y'_1 = \frac{1}{12h}(-3y_0 - 10y_1 + 18y_2 - 6y_3 + y_4) - \frac{h^4}{20} y^{(5)}(\xi),$$

$$y'_2 = \frac{1}{12h}(y_0 - 8y_1 + 8y_3 - y_4) + \frac{h^4}{30} y^{(5)}(\xi),$$

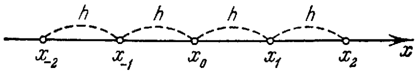
$$y'_3 = \frac{1}{12h}(-y_0 + 6y_1 - 18y_2 + 10y_3 + 3y_4) - \frac{h^4}{20} y^{(5)}(\xi),$$

$$y'_4 = \frac{1}{12h}(3y_0 - 16y_1 + 36y_2 - 48y_3 + 25y_4) + \frac{h^4}{4} y^{(5)}(\xi)$$

An examination of formulas I to III shows that if the number of points is odd and the derivative is taken at the midpoint, then the corresponding formula for numerical differentiation has a simpler expression and higher accuracy.

Below we give the formulas (for $n = 2$ and $n = 4$) for such *central derivatives* [3]; in order to exhibit the symmetry, we have changed the numbering of the points (Fig. 66):

Fig. 66



I. $n = 2$.

$$y'_0 = \frac{1}{2h}(y_1 - y_{-1}) - \frac{h^2}{6} y^{(3)}(\xi)$$

where $y_i = y(x_i)$ and $i = -1, 0, 1$.

II. $n = 4$.

$$y'_0 = \frac{2}{3h}(y_1 - y_{-1}) - \frac{1}{12h}(y_2 - y_{-2}) + \frac{h^4}{30} y^{(5)}\xi$$

where $y_i = y(x_i)$ and $i = -2, -1, 0, 1, 2$.

15.5 GRAPHICAL DIFFERENTIATION

The problem of graphical differentiation consists in constructing the graph of the derivative

$$Y = f'(x)$$

of a function $y = f(x)$ on the basis of the graph of that function.

We start with the graph of the function $y=f(x)$ (Fig. 67). To construct (to a known scale l) the graph of its derivative, choose on the given curve a sufficiently dense set of points 1, 2, 3, 4, 5,... including, if possible, the characteristic points of the graph. "By eye" construct tangents to the graph of the function at these points. Then, on the x-axis choose the point $P(-l, 0)$ (pole) and draw straight lines parallel to the corresponding tangents $P1'$, $P2'$, $P3'$, $P4'$, $P5'$, ... to their intersection with the y-axis. The segments of the y-axis $01'$, $02'$, $03'$, $04'$, $05'$, ... are, respectively, magnitudes proportional to the values of the derivative $y'=f'(x)$ at the chosen points; that is to say, they are ordinates of the graph of the derivative. Indeed, to take an example, for point 1 in Fig. 67 we have

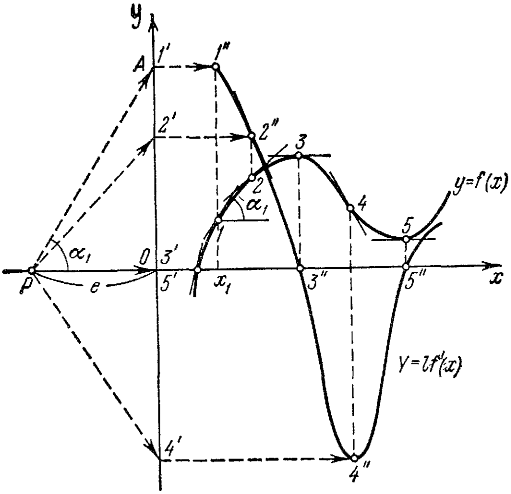


Fig. 67

$$OA = l \tan \alpha_1 = lf'(x_1)$$

We obtain similar results for all other points. Therefore, the points $1''$, $2''$, $3''$, $4''$, $5''$, ... of intersection of the parallels passing through the points $1'$, $2'$, $3'$, $4'$, $5'$, ... with the corresponding vertical lines passing through the points of tangency 1, 2, 3, 4, 5, ... belong to the graph of the derivative $y=lf'(x)$.

By joining the points $1''$, $2''$, $3''$, $4''$, $5''$, ... with a line the nature of which has regard for the positions of the intermediate points, we obtain an approximate graph of the derivative y' to the scale l. If we choose $l=1$, the graph of the derivative is then drawn to full scale.

To increase the accuracy of the graphical construction, it is advisable first to determine the direction of the tangent and then to give an indication of the point of tangency. To do this, subdivide the graph of the function into small segments that differ but slightly from straight-line segments. Let us consider one of them, AB in Fig. 68. Construct a family of chords parallel to the secant AB. The locus of the midpoints of these chords is the curve K intersecting the graph of the function at C, the tangent at C being parallel to the secant AB. Using this technique, we can find the point and the corresponding direction of the tangent on each line segment. The subsequent construction is carried out as indicated above.

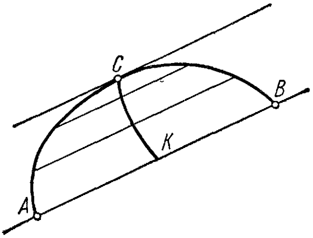


Fig. 68

More detailed descriptions may be found in the special literature (see, for example, [5]).

*15.6 ON THE APPROXIMATE CALCULATION OF PARTIAL DERIVATIVES

If a function $z=f(x,y)$ is given on a rectangular grid

$$x=x_0+ih, \quad y=y_0+jk$$

$(i,\ j=0,\ 1,\ 2,\ \ldots)$, then it may be represented in approximate fashion by the interpolation formula (Sec. 14.26)

$$\begin{aligned} z = z_{00} &+ [p\Delta^{1+0}z_{00}+q\Delta^{0+1}z_{00}]+ \\ &+\frac{1}{2!}[p(p-1)\Delta^{2+00}z_{00}+2pq\,\Delta^{1+1}z_{00}+q(q-1)\,\Delta^{0+2}z_{00}]+ \\ &+\frac{1}{3!}[p(p-1)(p-2)\,\Delta^{3+0}z_{00}+3p(p-1)\,q\Delta^{2+1}z_{00}+ \\ &+3pq(q-1)\,\Delta^{1+2}z_{00}+q(q-1)(q-2)\,\Delta^{0+3}z_{00}]+\ldots \quad (1) \end{aligned}$$

where

$$p=\frac{x-x_0}{h}, \qquad q=\frac{y-y_0}{k}$$

and $\Delta^{m+n}z_{00}=\Delta^{m+n}_{x^m y^n}z\,(0,\ 0)$ are mixed double differences.

From formula (1) it is easy to find the partial derivatives

$$\frac{\partial z}{\partial x}=\frac{\partial x}{\partial p}\cdot\frac{dp}{dx}=\frac{1}{h}\frac{\partial z}{\partial p},\quad \frac{\partial z}{\partial y}=\frac{\partial z}{\partial q}\cdot\frac{dq}{dy}=\frac{1}{k}\frac{\partial z}{\partial q}$$

and so forth.

REFERENCES FOR CHAPTER 15

[1] *A. N. Krylov, Lectures on Approximate Computations,* 1954, 6th edition, p. 228 (in Russian).
[2] *James B. Scarborough, Numerical Mathematical Analysis,* 1955, Chapter VII.
[3] *William E. Milne, Numerical Calculus,* 1949, Chapter IV.
[4] *Sh. E. Mikeladze, Numerical Methods of Mathematical Analysis,* 1953, Chapter XII (in Russian).
[5] *Carl Runge, Graphische Methoden,* 1919, III Kapitel, § 14.

Chapter 16

APPROXIMATE INTEGRATION OF FUNCTIONS

16.1 GENERAL REMARKS

If a function $f(x)$ is continuous on an interval $[a, b]$ and its antiderivative $F(x)$ is known, then the definite integral of this function from a to b may be computed from the *Newton-Leibniz formula*

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{1}$$

where $F'(x) = f(x)$.

However, in many cases the antiderivative $F(x)$ cannot be found by elementary means or is too involved; as a result, computation of the definite integral by formula (1) may be difficult or practically impossible.

Moreover, in practical situations, the integrand $f(x)$ is often specified in tabular form and then the whole concept of an antiderivative is meaningless. Similar problems arise in the computation of multiple integrals. Therefrom stems the great importance of approximate, primarily *numerical, methods* for computing definite integrals.

The problem of the numerical integration of a function consists in computing the value of a definite integral on the basis of a series of values of the integrand.

The numerical computation of a single integral is called *mechanical quadrature*, that of a double integral, *mechanical cubature*. We will call the respective formulas, *quadrature* and *cubature formulas*.

Let us first examine the numerical computation of single integrals. The ordinary technique of mechanical quadrature consists in replacing the given function $f(x)$ on the interval $[a, b]$ under consideration by an interpolating or approximating function $\varphi(x)$ of a simple kind (say a polynomial) and approximately setting

$$\int_a^b f(x)\,dx = \int_a^b \varphi(x)\,dx \tag{2}$$

The function $\varphi(x)$ must be such that the integral $\int_a^b \varphi(x)\,dx$ can be evaluated directly.

If the function $f(x)$ is given analytically, then the question is posed of estimating the error in formula (2).

Let us examine in more detail the use, for this purpose, of the Lagrange interpolation polynomial (Sec. 14.12).

Suppose, for the function $y=f(x)$, we know the corresponding values at $n+1$ points $x_0, x_1, x_2, \ldots, x_n$ of $[a, b]$:

$$f(x_i)=y_i \quad (i=0,\ 1,\ 2,\ \ldots,\ n) \tag{3}$$

It is required to find approximately

$$\int_a^b y\,dx=\int_a^b f(x)\,dx$$

Using the given values y_i, construct the Lagrange polynomial

$$L_n(x)=\sum_{i=0}^{n}\frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,y_i \tag{4}$$

where

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ldots(x-x_n)$$

and

$$L_n(x_i)=y_i \quad (i=0,\ 1,\ 2,\ \ldots,\ n)$$

Replacing the function $f(x)$ by the polynomial $L_n(x)$, we get

$$\int_a^b f(x)\,dx=\int_a^b L_n(x)\,dx+R_n[f] \tag{5}$$

where $R_n[f]$ is the error in the quadrature formula (5) *(remainder term)*. From this, using (4), we get the approximate quadrature formula

$$\int_a^b y\,dx=\sum_{i=0}^{n} A_i y_i \tag{6}$$

where

$$A_i=\int_a^b \frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,dx \quad (i=0,\ 1,\ 2,\ \ldots,\ n) \tag{7}$$

If the limits of integration a and b are interpolation points, then the quadrature formula (6) is of the "*closed type*", otherwise it is of the "*open type*".

With respect to computation of the coefficients A_i, note that

(1) the coefficients A_i are independent of the choice of the function $f(x)$ for a given arrangement of the points;

(2) for a polynomial of degree n, formula (6) is exact because in that case $L_n(x) \equiv f(x)$; hence, in particular, formula (6) is exact for $y = x^k$ $(k = 0, 1, \ldots, n)$; that is, $R_n[x^k] = 0$ for $k = 0, 1, \ldots, n$.

Putting $y = x^k$ $(k = 0, 1, 2, \ldots, n)$ in (6), we get a linear system of $n+1$ equations:

$$\left.\begin{aligned} I_0 &= \sum_{i=0}^{n} A_i, \\ I_1 &= \sum_{i=0}^{n} A_i x_i, \\ &\cdots\cdots\cdots \\ I_n &= \sum_{i=0}^{n} A_i x_i^n \end{aligned}\right\} \tag{8}$$

where

$$I_k = \int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{k+1} \qquad (k = 0, 1, \ldots, n)$$

from which it is possible to determine the coefficients $A_0, A_1, \ldots, A_n$ [1], [2]. The determinant of system (8) is the Vandermonde determinant

$$D = \prod_{i>j} (x_i - x_j) \neq 0$$

Note that in the application of this method the actual construction of the Lagrange polynomial $L_n(x)$ is unnecessary.

A simple method for computing the errors of quadrature formulas has been worked out by S. M. Nikolsky [3].

Example. Derive a quadrature formula of the form

$$\int_0^1 y\,dx = A_0 y\left(\frac{1}{4}\right) + A_1 y\left(\frac{1}{2}\right) + A_2 y\left(\frac{3}{4}\right) \tag{9}$$

Solution. Putting

$$y = x^k \qquad (k = 0, 1, 2)$$

in (9) and noting that

$$\int_0^1 dx = 1, \qquad \int_0^1 x\,dx = \frac{1}{2}, \qquad \int_0^1 x^2\,dx = \frac{1}{3}$$

we get the system

$$\left.\begin{aligned} 1 &= A_0 + A_1 + A_2, \\ \frac{1}{2} &= \frac{1}{4}A_0 + \frac{1}{2}A_1 + \frac{3}{4}A_2, \\ \frac{1}{3} &= \frac{1}{16}A_0 + \frac{1}{4}A_1 + \frac{9}{16}A_2 \end{aligned}\right\}$$

whence

$$A_0 = \frac{2}{3}, \quad A_1 = -\frac{1}{3}, \quad A_2 = \frac{2}{3}$$

and, thus,

$$\int_0^1 y\,dx = \frac{2}{3}y\left(\frac{1}{4}\right) - \frac{1}{3}y\left(\frac{1}{2}\right) + \frac{2}{3}y\left(\frac{3}{4}\right) \tag{10}$$

The quadrature formula (10) is of the open type and is exact for all polynomials of degree not exceeding two. It is easy to see that (10) yields a correct result for $y = x^3$ as well, and so this formula is also exact for third-degree polynomials.

16.2 NEWTON-COTES QUADRATURE FORMULAS

Suppose, for a given function $y = f(x)$, it is required to compute the integral

$$\int_a^b y\,dx$$

Choosing a spacing

$$h = \frac{b-a}{n}$$

divide the interval $[a, b]$ by means of equally spaced points

$$x_0 = a, \ x_i = x_0 + ih \qquad (i = 1, 2, \ldots, n-1), \ x_n = b$$

into n equal parts, and let

$$y_i = f(x_i) \qquad (i = 0, 1, 2, \ldots, n)$$

Replacing the function y by an appropriate Lagrange interpolation polynomial $L_n(x)$, we obtain the approximate quadrature formula

$$\int_{x_0}^{x_n} y\,dx = \sum_{i=0}^{n} A_i y_i \tag{1}$$

where A_i are certain constant coefficients.

We derive explicit expressions for the coefficients A_i of formula (1). As is known (Sec. 14.12),

$$L_n(x)=\sum_{i=0}^{n} p_i(x)\,y_i \tag{2}$$

where

$$p_i(x)=\frac{(x-x_0)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)} \tag{3}$$

Introducing the notation

$$q=\frac{x-x_0}{h} \tag{4}$$

and

$$q^{[n+1]}=q(q-1)\dots(q-n) \tag{5}$$

we obtain [cf. formula (2) of Sec. 15.4]

$$L_n(x)=\sum_{i=0}^{n}\frac{(-1)^{n-i}}{i!\,(n-i)!}\cdot\frac{q^{[n+1]}}{q-i}\,y_i \tag{6}$$

Replacing in (1) the function y by the polynomial $L_n(x)$, we obtain, by virtue of (6),

$$A_i=\int_{x_0}^{x_n}\frac{(-1)^{n-i}}{i!\,(n-i)!}\cdot\frac{q^{[n+1]}}{q-i}\,dx$$

or, since

$$q=\frac{x-x_0}{h},\quad dq=\frac{dx}{h}$$

then, by a change of variables in the definite integral, we get

$$A_i=h\,\frac{(-1)^{n-i}}{i!\,(n-i)!}\int_0^n\frac{q^{[n+1]}}{q-i}\,dq \qquad (i=0,\ 1,\ 2,\ \dots,\ n)$$

Since

$$h=\frac{b-a}{n}$$

we ordinarily put

$$A_i=(b-a)\,H_i$$

where

$$H_i=\frac{1}{n}\,\frac{(-1)^{n-i}}{i!\,(n-i)!}\int_0^n\frac{q^{[n+1]}}{q-i}\,dq \qquad (i=0,\ 1,\ 2,\ \dots,\ n) \tag{7}$$

are constants called *Cotes coefficients* (see, for example, [1], [4]).

Then the quadrature formula (1) assumes the form

$$\int_a^b y\,dx = (b-a)\sum_{i=0}^{n} H_i y_i \tag{8}$$

where

$$h=\frac{b-a}{n} \quad\text{and}\quad y_i=f_i(a+ih) \qquad (i=0,\ 1,\ \ldots,\ n)$$

It is easy to see that the following relations are valid:

$$(1)\ \sum_{i=0}^{n} H_i = 1, \qquad (2)\ H_i = H_{n-i}$$

16.3 THE TRAPEZOIDAL FORMULA AND ITS REMAINDER TERM

Applying formula (7) of the preceding section, we have, for $n=1$,

$$H_0 = -\int_0^1 \frac{q(q-1)}{q}\,dq = \frac{1}{2},$$

$$H_1 = \int_0^1 q\,dq = \frac{1}{2}$$

whence

$$\int_{x_0}^{x_1} y\,dx = \frac{h}{2}(y_0+y_1) \tag{1}$$

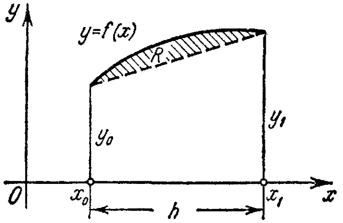


Fig. 69

We thus obtain the *trapezoidal formula* for approximate computation of a definite integral (Fig. 69).

The remainder term (error) of the quadrature formula (1) is

$$R = \int_{x_0}^{x_1} y\,dx - \frac{h}{2}(y_0+y_1)$$

Assuming that $y \in C^{(2)}[a,b]$, we derive a simple formula for the remainder term. We will regard $R=R(h)$ as a function of the spacing h; then we can put

$$R(h) = \int_{x_0}^{x_0+h} y\,dx - \frac{h}{2}[y(x_0)+y(x_0+h)]$$

Differentiating this formula with respect to h twice in succession, we obtain

$$R'(h)=y(x_0+h)-\frac{1}{2}[y(x_0)+y(x_0+h)]-\frac{h}{2}y'(x_0+h)=$$
$$=\frac{1}{2}[y(x_0+h)-y(x_0)]-\frac{h}{2}y'(x_0+h)$$

and

$$R''(h)=\frac{1}{2}y'(x_0+h)-\frac{1}{2}y'(x_0+h)-\frac{h}{2}y''(x_0+h)=$$
$$=-\frac{h}{2}y''(x_0+h)$$

Note that

$$R(0)=0,\ R'(0)=0$$

From this, integrating with respect to h and using the mean-value theorem, we successively derive

$$R'(h)=R'(0)+\int_0^h R''(t)\,dt=-\frac{1}{2}\int_0^h ty''(x_0+t)\,dt=$$
$$=-\frac{1}{2}y''(\xi_1)\int_0^h t\,dt=-\frac{h^2}{4}y''(\xi_1)$$

where $\xi_1\in(x_0,\ x_0+h)$ and

$$R(h)=R(0)+\int_0^h R'(t)\,dt=-\frac{1}{4}\int_0^h t^2y''(\xi_1)\,dt=$$
$$=-\frac{1}{4}y''(\xi)\int_0^h t^2dt=-\frac{h^3}{12}y''(\xi)$$

where $\xi\in(x_0,\ x_0+h)$.

Thus, we finally have

$$R=-\frac{h^3}{12}y''(\xi)$$

where $\xi\in(x_0,\ x_1)$.

From the foregoing, it follows, in particular, that if $y''>0$, then formula (1) yields the value of the integral with an *excess*, but if $y''<0$, it yields the value with a *deficit*.

16.4 SIMPSON'S FORMULA AND ITS REMAINDER TERM

From formula (7) of Sec. 16.2 we get, for $n=2$,

$$H_0=\frac{1}{2}\cdot\frac{1}{2}\int_0^2(q-1)(q-2)\,dq=\frac{1}{4}\left(\frac{8}{3}-6+4\right)=\frac{1}{6},$$

$$H_1 = -\frac{1}{2} \cdot \frac{1}{1} \int_0^2 q(q-2)\,dq = \frac{2}{3},$$

$$H_2 = \frac{1}{2} \cdot \frac{1}{2} \int_0^2 q(q-1)\,dq = \frac{1}{6}$$

Hence, since $x_2 - x_0 = 2h$, we have

$$\int_{x_0}^{x_2} y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2) \tag{1}$$

which is the formula of *Simpson's rule*. Geometrically, this formula is obtained by replacing the given curve $y = f(x)$ by the parabola $y = L_2(x)$ passing through three points $M_0(x_0, y_0)$, $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$ (Fig. 70).

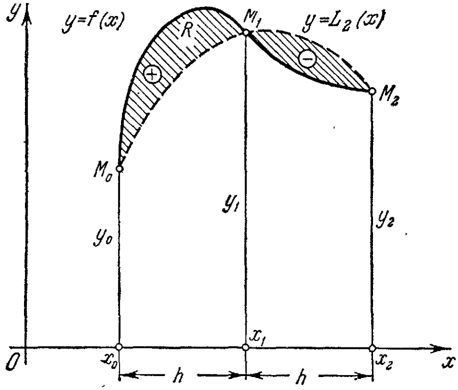


Fig. 70

The remainder term of Simpson's formula is

$$R = \int_{x_0}^{x_2} y\,dx - \frac{h}{3}(y_0 + 4y_1 + y_2)$$

Assuming that $y \in C^{(4)}[a, b]$, we derive a simpler expression for R in a manner similar to what was done for the trapezoidal formula. Fixing the midpoint x_1 and regarding $R = R(h)$ as a function of the spacing $h\,(h \geqslant 0)$, we get

$$R(h) = \int_{x_1 - h}^{x_1 + h} y\,dx - \frac{h}{3}\,[y(x_1 - h) + 4y(x_1) + y(x_1 + h)]$$

whence, differentiating the function $R(h)$ three times in succession with respect to h, we obtain

$$R'(h)=[y(x_1+h)+y(x_1-h)]-\frac{1}{3}[y(x_1-h)+4y(x_1)+y(x_1+h)]-$$

$$-\frac{h}{3}[-y'(x_1-h)+y'(x_1+h)]=\frac{2}{3}[y(x_1-h)+y(x_1+h)]-$$

$$-\frac{4}{3}y(x_1)-\frac{h}{3}[-y'(x_1-h)+y'(x_1+h)],$$

$$R''(h)=\frac{2}{3}[-y'(x_1-h)+y'(x_1+h)]-$$

$$-\frac{1}{3}[-y'(x_1-h)+y'(x_1+h)]-\frac{h}{3}[y''(x_1-h)+y''(x_1+h)]=$$

$$=\frac{1}{3}[-y'(x_1-h)+y'(x_1+h)]-\frac{h}{3}[y''(x_1-h)+y''(x_1+h)],$$

$$R'''(h)=\frac{1}{3}[y''(x_1-h)+y''(x_1+h)]-$$

$$-\frac{1}{3}[y''(x_1-h)+y''(x_1+h)]-\frac{h}{3}[-y'''(x_1-h)+y'''(x_1+h)]=$$

$$=-\frac{h}{3}[y'''(x_1+h)-y'''(x_1-h)]=-\frac{2h^2}{3}y^{\mathrm{IV}}(\xi_3)$$

where $\xi_3\in(x_1-h,\ x_1+h)$.

Moreover, we have

$$R(0)=0,\ R'(0)=0,\ R''(0)=0$$

Successively integrating $R'''(h)$ and using the mean-value theorem, we find

$$R''(h)=R''(0)+\int_0^h R'''(t)\,dt=-\frac{2}{3}\int_0^h t^2y^{\mathrm{IV}}(\xi_3)\,dt=$$

$$=-\frac{2}{3}y^{\mathrm{IV}}(\xi_2)\int_0^h t^2\,dt=-\frac{2}{9}h^3y^{\mathrm{IV}}(\xi_2)$$

where $\xi_2\in(x_1-h,\ x_1+h)$,

$$R'(h)=R'(0)+\int_0^h R''(t)\,dt=-\frac{2}{9}\int_0^h t^3y^{\mathrm{IV}}(\xi_2)\,dt=$$

$$=-\frac{2}{9}y^{\mathrm{IV}}(\xi_1)\int_0^h t^3\,dt=-\frac{1}{18}h^4y^{\mathrm{IV}}(\xi_1)$$

where $\xi_1 \in (x_1 - h,\ x_1 + h)$,

$$R(h) = R(0) + \int_0^h R'(t)\,dt = -\frac{1}{18}\int_0^h t^4 y^{\mathrm{IV}}(\xi_1)\,dt =$$

$$= -\frac{1}{18} y^{\mathrm{IV}}(\xi) \int_0^h t^4\,dt = -\frac{h^5}{90} y^{\mathrm{IV}}(\xi)$$

where $\xi \in (x_1 - h,\ x_1 + h)$.

Thus, the *remainder term of Simpson's formula* is

$$R = -\frac{h^5}{90} y^{\mathrm{IV}}(\xi) \tag{2}$$

where $\xi \in (x_0,\ x_2)$.

Hence, this formula is *exact* for polynomials not only of degree two but of degree three as well, which means that Simpson's rule has a rather high accuracy for a relatively small number of ordinates.

16.5 NEWTON-COTES FORMULAS OF HIGHER ORDERS

Carrying out the appropriate computations for $n=3$, we obtain from formula (7) of Sec. 16.2 *Newton's quadrature formula*

$$\int_{x_0}^{x_3} y\,dx = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3) \tag{1}$$

(*three-eighths rule*).

The remainder term of formula (1) is [2]

$$R = -\frac{3h^5}{80} y^{\mathrm{IV}}(\xi)$$

where $\xi \in (x_0,\ x_3)$.

The subsequent Newton-Cotes quadrature formulas are given in [1], [2]. The remainder terms of these formulas are given by Steffensen (see [1], [5], [6]).

Note that for sufficient smoothness of the function $y = f(x)$ the error in a Newton-Cotes formula employing $n+1$ ordinates is at least of the order of [1], [6]

$$R = O\left[h^{2E\left(\frac{n}{2}\right)+3}\right]$$

where $E\left(\frac{n}{2}\right)$ is the largest integer in the fraction $\frac{n}{2}$.

From this we see that quadrature formulas employing an odd number of ordinates are more advantageous as to the degree of accuracy. See the table of Cotes coefficients (Table 65). For the sake of notational convenience, the Cotes coefficients for each n

TABLE 65
COTES COEFFICIENTS

n	$\hat{H}_0$	$\hat{H}_1$	$\hat{H}_2$	$\hat{H}_3$	$\hat{H}_4$	$\hat{H}_5$	$\hat{H}_6$	$\hat{H}_7$	$\hat{H}_8$	Common denominator N
1	1	1								2
2	1	4	1							6
3	1	3	3	1						8
4	7	32	12	32	7					90
5	19	75	50	50	75	19				288
6	41	216	27	272	27	216	41			840
7	751	3577	1323	2989	2989	1323	3577	751		17280
8	989	5888	—928	10496	—4540	10496	—928	5888	989	28350

are given in the form of fractions:

$$H_i=\frac{\hat{H}_i}{N}$$

with common denominator N. Note as a check that

$$\sum_{i=0}^{n} \hat{H}_i=N$$

It is worth bearing in mind that the Cotes coefficients for large n may be negative (see, for instance, $n=8$).

Example. Evaluate

$$I=\int_0^1 \frac{dx}{1+x}$$

using a Newton-Cotes formula employing seven ordinates ($n=6$).

Solution. Taking a spacing of

$$h=\frac{1-0}{6}=\frac{1}{6}$$

we form a table of values (see Table 66) where for convenience we assume $\hat{H}_i=840\,H_i$.
Whence

$$I=\frac{1}{840}\cdot 581.994372=0.6933$$

TABLE 66
EVALUATING AN INTEGRAL BY THE NEWTON-COTES FORMULA

i	x_i	y_i	$\hat{H}_i$	$\hat{H}_i y_i$
0	0	1	41	41
1	$\frac{1}{6}$	$\frac{6}{7}$	216	185.142857
2	$\frac{1}{3}$	$\frac{3}{4}$	27	20.25
3	$\frac{1}{2}$	$\frac{2}{3}$	272	181.333333
4	$\frac{2}{3}$	$\frac{3}{5}$	27	16.2
5	$\frac{5}{6}$	$\frac{6}{11}$	216	117.818182
6	1	$\frac{1}{2}$	41	20.25
$\sum$				581.994372

The exact value is

$$I = \ln 2 = 0.69315 \ldots$$

Since the Cotes coefficients are extremely involved for a large number of ordinates, a practical procedure for approximating definite integrals is as follows: the interval of integration is divided into a sufficiently large number of subintervals to each of which is applied a Newton-Cotes quadrature formula employing a small number of ordinates (see, for instance, [7]). Formulas of simpler structure are then obtained and the accuracy of these formulas may be arbitrarily high.

We will now consider some examples of such formulas.

16.6 GENERAL TRAPEZOIDAL FORMULA (TRAPEZOIDAL RULE)

To evaluate the integral

$$\int_a^b y\,dx$$

divide the interval of integration $[a, b]$ into n equal parts $[x_0, x_1]$, $[x_1, x_2]$, $\ldots$, $[x_{n-1}, x_n]$ and to each apply the trapezoidal rule [see Sec. 16.3, formula (1)]. Setting $h=\frac{b-a}{n}$ and denoting by $y_i=f(x_i)$ $(i=0, 1, \ldots, n)$ the values of the integrand at the points x_i, we have

$$\int_a^b y\,dx=\frac{h}{2}(y_0+y_1)+\frac{h}{2}(y_1+y_2)+\ldots+\frac{h}{2}(y_{n-1}+y_n)$$

or

$$\int_a^b y\,dx=h\left(\frac{y_0}{2}+y_1+y_2+\ldots+y_{n-2}+y_{n-1}+\frac{y_n}{2}\right) \tag{1}$$

Geometrically, formula (1) is obtained by replacing the graph of the integrand function $y=f(x)$ by a polygonal line (Fig. 71).

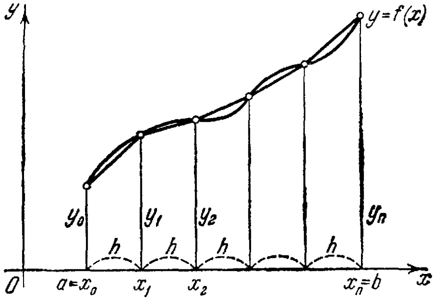


Fig. 71

If $y\in C^{(2)}[a, b]$, then the *remainder term* of the quadrature formula (1) is, by (2) of Sec. 16.3,

$$R=\int_{x_0}^{x_n} y\,dx-\frac{h}{2}\sum_{i=1}^{n}(y_{i-1}+y_i)=$$

$$=\sum_{i=1}^{n}\left[\int_{x_{i-1}}^{x_i} y\,dx-\frac{h}{2}(y_{i-1}+y_i)\right]=-\frac{h^3}{12}\sum_{i=1}^{n}y''(\xi_i) \tag{2}$$

where $\xi_i\in(x_{i-1}, x_i)$.

Consider the arithmetic mean

$$\mu=\frac{1}{n}\sum_{i=1}^{n}y''(\xi_i). \tag{3}$$

Clearly, μ lies between the smallest value m_2 and the largest value M_2 of the second derivative y'' on the interval $[a, b]$. Thus

$$m_2 \leqslant \mu \leqslant M_2$$

Since y'' is continuous on $[a, b]$, it assumes all intermediate values between m_2 and M_2. A point $\xi \in [a, b]$ therefore exists such that

$$\mu = f''(\xi)$$

From formulas (2) and (3) we have

$$R = -\frac{nh^3}{12} y''(\xi) = -\frac{(b-a)h^2}{12} y''(\xi)$$

where $\xi \in [a, b]$.

16.7 SIMPSON'S GENERAL FORMULA (PARABOLIC RULE)

Let $n = 2m$ be an even number and let $y_i = f(x_i)$ $(i = 0, 1, 2, \ldots, n)$ be the values of the function $y = f(x)$ for equally spaced points $a = x_0, x_1, \ldots, x_n = b$ with spacing

$$h = \frac{b-a}{n} = \frac{b-a}{2m}$$

Applying Simpson's rule [Sec. 16.4, formula (1)] to each doubled interval $[x_0, x_2]$, $[x_2, x_4]$, $\ldots$, $[x_{2m-2}, x_{2m}]$ of length $2h$ (Fig. 72), we get

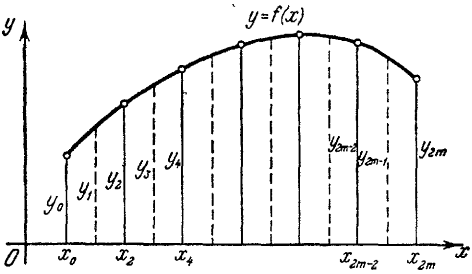


Fig. 72

$$\int_a^b y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2) + \frac{h}{3}(y_2 + 4y_3 + y_4) + \ldots$$

$$\ldots + \frac{h}{3}(y_{2m-2} + 4y_{2m-1} + y_{2m})$$

whence we obtain *Simpson's general formula:*

$$\int_a^b y\,dx = \frac{h}{3}[(y_0+y_{2m})+4(y_1+y_3+\ldots+y_{2m-1})+$$
$$+2(y_2+y_4+\ldots+y_{2m-2})] \quad (1)$$

Introducing the notation

$$\sigma_1 = y_1+y_3+\ldots+y_{2m-1},$$
$$\sigma_2 = y_2+y_4+\ldots+y_{2m}$$

we can write (1) more simply as

$$\int_a^b y\,dx = \frac{h}{3}[(y_0+y_n)+4\sigma_1+2\sigma_2] \quad (1')$$

If $y \in C^{(4)}[a, b]$, then the error in Simpson's formula on every doubled interval $[x_{2k-2}, x_{2k}]$ $(k=1, 2, \ldots, m)$ is given, on the basis of formula (2) of Sec. 16.4, by the formula

$$r_k = -\frac{h^5}{90}y^{\text{IV}}(\xi_k)$$

where $\xi_k \in (x_{2k-2}, x_{2k})$. Summing all these errors, we get the *remainder term for Simpson's general formula* in the form

$$R = -\frac{h^5}{90}\sum_{k=1}^{m} y^{\text{IV}}(\xi_k)$$

Since $y^{\text{IV}}(x)$ is continuous on $[a, b]$ there is a point $\xi \in [a, b]$ such that

$$y^{\text{IV}}(\xi) = \frac{1}{m}\sum_{k=1}^{m} y^{\text{IV}}(\xi_k)$$

We therefore have

$$R = -\frac{mh^5}{90}y^{\text{IV}}(\xi) = -\frac{(b-a)h^4}{180}y^{\text{IV}}(\xi) \quad (2)$$

where $\xi \in [a, b]$.

If the maximum admissible error $\varepsilon > 0$ is given, then, denoting

$$M_4 = \max|y^{\text{VI}}(x)|,$$

we will have, for determining the spacing h, the inequality

$$(b-a)\frac{h^4}{180}M_4 < \varepsilon$$

whence

$$h < \sqrt[4]{\frac{180\varepsilon}{(b-a)M_4}}$$

Thus, h is of the order of $\sqrt[4]{\varepsilon}$.

In many cases, it is extremely difficult to estimate the error in Simpson's quadrature formula (1) using (2). Then a double computation is carried out with spacings h and $2h$, and it is taken that the coincident decimals belong to the exact value of the integral.

There is another technique of practical convenience for computing the error in Simpson's quadrature formula. Assuming that the derivative $y^{\mathrm{IV}}(x)$ varies slowly on the interval $[a, b]$ we obtain by (2) an approximate expression for the desired error

$$R = Mh^4$$

where the coefficient M will be considered constant. Let Σ_h and Σ_H be approximate values of the integral

$$I = \int_a^b y\,dx$$

obtained by Simpson's rule with spacing h and $H = 2h$, respectively. We have

$$I = \Sigma_h + Mh^4$$

and

$$I = \Sigma_H + M(2h)^4$$

whence

$$R = \frac{\Sigma_h - \Sigma_H}{15}$$

For the approximate value of the integral I it is advisable to take the corrected value

$$I = \Sigma_h + \frac{\Sigma_h - \Sigma_H}{15}$$

Note that if the number of divisions n is a multiple of 4, then the sum Σ_H may be computed by using tabular values, taking every other one.

Example. Using Simpson's formula, compute the integral

$$I = \int_0^1 \frac{dx}{1+x}$$

taking $n = 10$.

Solution. We have $2m = 10$, whence

$$h = \frac{1-0}{10} = 0.1$$

The results of the computations are given in Table 67.

TABLE 67

EVALUATING AN INTEGRAL BY SIMPSON FORMULA

i	x_i	y_{2j-1}	y_{2j}
0	0		$\boxed{y_0 = 1,00000}$
1	0.1	0,90909	
2	0.2		0.83333
3	0.3	0,76923	
4	0.4		0.71429
5	0.5	0,66667	
6	0.6		0.62500
7	0.7	0.58824	
8	0.8		0.55556
9	0.9	0.52632	
10	1.0		$\boxed{0.50000 = y_n}$
Σ		$\boxed{3.45955}$ (σ_1)	$\boxed{2.72818}$ (σ_2)

By formula (1′) we have

$$I \approx \frac{h}{3}(y_0 + y_n + 4\sigma_1 + 2\sigma_2) = 0.69315 \qquad (3)$$

Let us compute the error of the result (3). The total error R is made up of the error R_1 of operation and the remainder term R_2. Clearly,

$$R_1 = \sum_{i=0}^{n} A_i \varepsilon$$

where A_i are the coefficients in Simpson's formula and ε is the maximal rounding error for the values of the integrand.

In our case

$$R_1 = nh\varepsilon = (b-a)\,\varepsilon = 1 \cdot \frac{1}{2} \cdot 10^{-5} = 0.5 \cdot 10^{-5}$$

We estimate the remainder term by formula (2). Since

$$y = \frac{1}{1+x} = (1+x)^{-1}$$

hen

$$y^{\text{IV}} = (-1)(-2)(-3)(-4)(1+x)-5 = \frac{24}{(1+x)^5}$$

whence

$$\max |y^{\text{IV}}| = 24 \text{ for } 0 \leqslant x \leqslant 1$$

and, therefore,

$$|R_2| \leqslant 1 \cdot \frac{(0.1)^4}{180} \cdot 24 = 1.3 \cdot 10^{-5}$$

Thus, the limiting total error is

$$R = 0.5 \cdot 10^{-5} + 1.3 \cdot 10^{-5} = 1.8 \cdot 10^{-5} < 0.00002$$

and, hence,

$$I \doteq 0.69315 \pm 0.00002$$

16.8 ON CHEBYSHEV'S QUADRATURE FORMULA

Let us consider the quadrature formula

$$\int_{-1}^{1} f(t)\,dt = \sum_{i=1}^{n} B_i f(t_i) \tag{1}$$

where B_i are constant coefficients.

Chebyshev suggested choosing the abscissas t_i so that:

(1) the coefficients B_i are equal,

(2) he quadrature formula (1) is exact for all polynomials of degree up to n inclusive.

Let us show how the B_i and t_i can then be found. Setting

$$B_1 = B_2 = \ldots = B_n = B$$

and noting that for $f(t) \equiv 1$, we have

$$2 = \sum_{i=1}^{n} B_i$$

whence we obtain

$$B = \frac{2}{n}$$

Consequently, *Chebyshev's quadrature formula* is of the form

$$\int_{-1}^{1} f(t)\,dt = \frac{2}{n} \sum_{i=1}^{n} f(t_i) \tag{2}$$

To determine the abscissas t_i, note that formula (2), by Condition (2), must be exact for functions of the form

$$f(t)=t,\ t^2,\ \ldots,\ t^n$$

Substituting these functions into (2), we obtain the system of equations

$$\left.\begin{aligned}
&t_1+t_2+\ldots+t_n=0,\\
&t_1^2+t_2^2+\ldots+t_n^2=\frac{n}{3},\\
&t_1^3+t_2^3+\ldots+t_n^3=0,\\
&t_1^4+t_2^4+\ldots+t_n^4=\frac{n}{5},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&t_1^n+t_2^n+\ldots+t_n^n=\frac{n\,[1-(-1)^{n+1}]}{2\,(n+1)}
\end{aligned}\right\}\qquad(3)$$

from which we can determine the unknowns t_i ($i=1,\ 2,\ \ldots,\ n$). Chebyshev demonstrated that the solution of system (3) reduces to finding the roots of a certain algebraic equation of degree n [6], [8]. Table 68 lists the values of the roots t_i of system (3) for $n=2,\ 3,\ \ldots,\ 7$.

TABLE 68

VALUES OF ABSCISSAS t_i IN CHEBYSHEV'S FORMULA

n	i	t_i	n	i	t_i
2	1; 2	∓0.577350	6	1; 6	∓0.866247
3	1; 3	∓0.707107		2; 5	∓0.422519
	2	0		3; 4	∓0.266635
4	1; 4	∓0.794654	7	1; 7	∓0.883862
	2; 3	∓0.187592		2; 6	∓0.529657
5	1; 5	∓0.832498		3; 5	∓0.323912
	2; 4	∓0.374541		4	0
	3	0			

As S. N. Bernstein demonstrated, the system (3) for $n=8$ and $n\geqslant 10$ does not have any real solutions. Therein lies the fundamental defect of Chebyshev's quadrature formula.

Exampel 1. Derive Chebyshev's formula with three ordinates ($n=3$).

Solution. We have the following system of equations for determining the abscissas t_i $(i=1, 2, 3)$:

$$\left.\begin{aligned} t_1+t_2+t_3&=0,\\ t_1^2+t_2^2+t_3^2&=1,\\ t_1^3+t_2^3+t_3^3&=0 \end{aligned}\right\} \tag{4}$$

Let us consider the symmetric functions of the roots

$$\begin{aligned} C_1&=t_1+t_2+t_3,\\ C_2&=t_1t_2+t_1t_3+t_2t_3,\\ C_3&=t_1t_2t_3 \end{aligned}$$

From system (4) we have

$$C_1=0,$$

$$C_2=\frac{1}{2}\left[(t_1+t_2+t_3)^2-(t_1^2+t_2^2+t_3^2)\right]=\frac{1}{2}(0-1)=-\frac{1}{2},$$

$$C_3=\frac{1}{6}\left[(t_1+t_2+t_3)^3-3(t_1+t_2+t_3)(t_1^2+t_2^2+t_3^2)+\right.$$

$$\left.+2(t_1^3+t_2^3+t_3^3)\right]=\frac{1}{6}(0-0+0)=0$$

From this we conclude that t_i are the roots of the auxiliary equation

$$t^3-C_1t^2+C_2t-C_3=0$$

or

$$t^3-\frac{1}{2}t=0$$

Hence, we can take

$$t_1=-\frac{\sqrt{2}}{2},\ t_2=0,\ t_3=\frac{\sqrt{2}}{2}$$

Thus, the corresponding Chebyshev formula has the form

$$\int_{-1}^{1} f(t)\,dt=\frac{2}{3}\left[f\left(-\frac{1}{\sqrt{2}}\right)+f(0)+f\left(\frac{1}{\sqrt{2}}\right)\right]$$

To apply the Chebyshev quadrature formula to an integral of the form

$$\int_a^b f(x)\,dx$$

it is necessary to transform it by the substitution

$$x=\frac{b+a}{2}+\frac{b-a}{2}t$$

which carries the interval $a \leqslant x \leqslant b$ into the interval $-1 \leqslant t \leqslant 1$ Applying Chebyshev's formula (2) to the transformed integral, we get

$$\int_a^b f(x)\,dx = \frac{b-a}{n} \sum_{i=1}^{n} f(x_i) \tag{5}$$

where

$$x_i = \frac{b+a}{2} + \frac{b-a}{2} t_i \tag{6}$$

and $t_i (i = 1, 2, \ldots, n)$ are the roots of system (3) (given in Table 68).

Chebyshev's quadrature formula is mostly used in shipbuilding.

Example 2. Evaluate the integral

$$I = \int_0^1 \frac{x\,dx}{1+x}$$

using Chebyshev's formula with five ordinates $(n = 5)$.

Solution. Introducing the notation

$$f(x) = \frac{x}{1+x}$$

we have

$$I = \frac{1}{5} [f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5)]$$

where, by formula (6),

$$x_1 = \frac{1}{2} + \frac{1}{2} t_1 = \frac{1}{2} + \frac{1}{2} \cdot (-0.83250) = 0.08375,$$
$$x_2 = \frac{1}{2} + \frac{1}{2} t_2 = \frac{1}{2} + \frac{1}{2} \cdot (-0.37454) = 0.31273,$$
$$x_3 = \frac{1}{2} + \frac{1}{2} t_3 = \frac{1}{2} + \frac{1}{2} \cdot 0 = 0.5$$
$$x_4 = 1 - x_2 = 0.68727$$
$$x_5 = 1 - x_1 = 0.91625$$

The respective values of $y_i = f(x_i)$ $(i = 1, 2, 3, 4, 5)$ of the integrand are listed in Table 69.

From this we get

$$I = \frac{1}{5} \cdot 1.5342 = 0.3068$$

By way of comparison, we give the exact value of the integral to six significant digits:

$$I = 0.306846 \ldots$$

TABLE 69

EVALUATING AN INTEGRAL BY CHEBYSHEV'S FORMULA

i	x_i	y_i
1	0.08375	0.0773
2	0.31273	0.2382
3	0.50000	0.3333
4	0.68727	0.4073
5	0.91625	0.4781
Σ		1.5342

16.9 GAUSSIAN QUADRATURE FORMULA

In this section we will need some facts about Legendre polynomials, which are polynomials of the form

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n] \qquad (n=0,\ 1,\ 2,\ \ldots)$$

The following are the basic properties of Legendre polynomials [1]:
(1) $P_n(1)=1,\ P_n(-1)=(-1)^n\ (n=0,\ 1,\ \ldots)$,

(2) $\int\limits_{-1}^{1} P_n(x)\cdot Q_k(x)\,dx=0 \qquad (k<n)$, where $Q_k(x)$ is any polynomial of degree k less than n;

(3) the Legendre polynomial $P_n(x)$ has n distinct real roots lying in the interval $(-1,\ 1)$.

Listed below are the first five Legendre polynomials (their graphs are shown in Fig. 73):

$$P_0(x)=1,$$
$$P_1(x)=x,$$
$$P_2(x)=\frac{1}{2}(3x^2-1),$$
$$P_3(x)=\frac{1}{2}(5x^3-3x),$$
$$P_4(x)=\frac{1}{8}(35x^4-30x^2+3)$$

Let us now derive the *Gaussian quadrature formula.* We first consider a function $y=f(t)$ specified on the standard interval $[-1,\ 1]$. The general case can readily be reduced to our case by a linear substitution of the independent variable.

We pose the problem as follows: how must one choose points $t_1, t_2, \ldots, t_n$ and coefficients $A_1, A_2, \ldots, A_n$ so that the quadrature formula

$$\int_{-1}^{1} f(t)\,dt = \sum_{i=1}^{n} A_i f(t_i) \tag{1}$$

is exact for all polynomials $f(t)$ of degree N as high as possible.

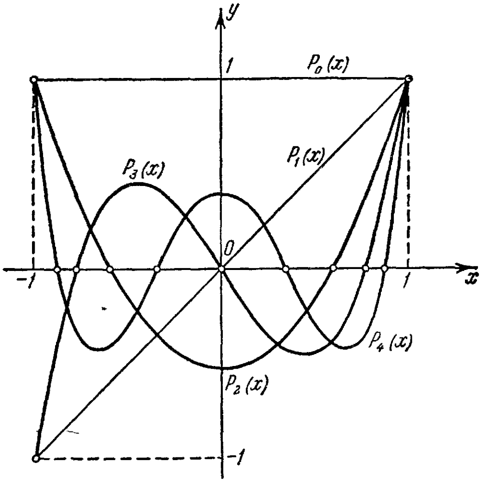


Fig. 73

Since we have at our disposal $2n$ constants t_i and A_i $(i = 1, 2, \ldots, n)$, and a polynomial of degree $2n-1$ is determined by $2n$ coefficients, this highest possible degree is in the general case clearly equal to $N = 2n-1$.

To ensure equation (1) it is necessary and sufficient that it be valid for

$$f(t) = 1,\ t,\ t^2,\ \ldots,\ t^{2n-1}$$

Indeed, setting

$$\int_{-1}^{1} t^k\,dt = \sum_{i=1}^{n} A_i t_i^k \qquad (k = 0, 1, 2, \ldots, 2n-1) \tag{2}$$

and

$$f(t) = \sum_{k=0}^{2n-1} C_k t^k$$

we get

$$\int_{-1}^{1} f(t)\,dt = \sum_{k=0}^{2n-1} C_k \int_{-1}^{1} t^k\,dt = \sum_{k=0}^{2n-1} C_k \sum_{i=1}^{n} A_i t_i^k =$$

$$= \sum_{i=1}^{n} A_i \sum_{k=0}^{2n-1} C_k t_i^k = \sum_{i=1}^{n} A_i f(t_i)$$

Thus, taking into account the relations

$$\int_{-1}^{1} t^k\,dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \dfrac{2}{k+1} & \text{for } k \text{ even} \\ 0 & \text{for } k \text{ odd} \end{cases}$$

we conclude that to solve the problem [2], [3], [6], it is sufficient to determine t_i and A_i from the following system of $2n$ equations:

$$\left.\begin{aligned} &\sum_{i=1}^{n} A_i = 2, \\ &\sum_{i=1}^{n} A_i t_i = 0, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ &\sum_{i=1}^{n} A_i t_i^{2n-2} = \frac{2}{2n-1}, \\ &\sum_{i=1}^{n} A_i t_i^{2n-1} = 0 \end{aligned}\right\} \tag{3}$$

The system (3) is a nonlinear system and its solution in the ordinary manner involves great mathematical difficulties. However, the following artificial device may be employed.

Consider the polynomials

$$f(t) = t^k P_n(t) \qquad (k = 0, 1, \ldots, n-1)$$

where $P_n(t)$ is Legendre's polynomial.

Since the degrees of these polynomials do not exceed $2n-1$, then formula (1) should, on the basis of (3), hold true, and

$$\int_{-1}^{1} t^k P_n(t)\,dt = \sum_{i=1}^{n} A_i t_i^k P_n(t_i) \qquad (k = 0, 1, \ldots, n-1) \tag{4}$$

On the other hand, by virtue of the orthogonality property of Legendre polynomials (Property 2), the equations

$$\int_{-1}^{1} t^k P_n(t)\,dt = 0 \quad \text{for} \quad k < n$$

are valid and therefore

$$\sum_{i=1}^{n} A_i t_i^k P_n(t_i) = 0 \qquad (k=0,\ 1,\ \dots,\ n-1) \tag{5}$$

Equations (5) will definitely be ensured for any values A_i if we put

$$P_n(t_i) = 0 \qquad (i=1,\ 2,\ \dots,\ n) \tag{6}$$

Thus, to achieve the maximum accuracy of the quadrature formula (1), it is sufficient to take for the points t_i the zeros of the respective Legendre polynomial. As is known (Property 3), these zeros are real and distinct and lie in the interval $(-1,\ 1)$. Knowing the abscissas of t_i, the coefficients $A_i (i=1,\ 2,\ \dots,\ n)$ can readily be found from the linear system of the first n equations of system (3). The determinant of this subsystem is the Vandermonde determinant

$$D = \prod_{i>j} (t_i - t_j) \neq 0$$

and, hence, A_i are determined unambiguously. It may be shown that the quadrature formula (1) with the thus obtained coefficients $A_i (i=1,\ 2,\ \dots,\ n)$ will be more exact for all polynomials of degree not higher than $2n-1$.

Formula (1), where the t_i are zeros of the Legendre polynomial $P_n(t)$ and the $A_i (i=1,\ 2,\ \dots,\ n)$ are determined from system (3), is called the *Gaussian quadrature formula*.

Example 1. Derive the Gaussian quadrature formula for the case of three ordinates ($n=3$).

Solution. The Legendre polynomial of degree three is

$$P_3(t) = \frac{1}{2}(5t^3 - 3t)$$

Equating this polynomial to zero, we find the roots

$$t_1 = -\sqrt{\frac{3}{5}} \approx -0.774597,$$
$$t_2 = 0,$$
$$t_3 = \sqrt{\frac{3}{5}} \approx 0.774597$$

To determine the coefficients, A_1, A_2, A_3, we have, by (3)

$$\left.\begin{aligned} A_1 + A_2 + A_3 &= 2, \\ -\sqrt{\frac{3}{5}} A_1 + \sqrt{\frac{3}{5}} A_3 &= 0, \\ \frac{3}{5} A_1 + \frac{3}{5} A_3 &= \frac{2}{3} \end{aligned}\right\}$$

whence

$$A_1 = A_3 = \frac{5}{9}, \qquad A_2 = \frac{8}{9}$$

Therefore,

$$\int_{-1}^{1} f(t)\,dt = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right]$$

TABLE 70

ELEMENTS OF THE GAUSSIAN FORMULA

n	i	t_i	A_i
1	1	0	2
2	1; 2	∓ 0.57735027	1
3	1; 3	∓ 0.77459667	$\frac{5}{9} = 0.55555556$
	2	0	$\frac{8}{9} = 0.88888889$
4	1; 4	∓ 0.86113631	0.34785484
	2; 3	∓ 0.33998104	0.65214516
5	1; 5	∓ 0.90617985	0.23692688
	2; 4	∓ 0.53846931	0.47862868
	3;	0	0.56888889
6	1; 6	∓ 0.93246951	0.17132450
	2; 5	∓ 0.66120939	0.36076158
	3; 4	∓ 0.23861919	0.46791394
7	1; 7	∓ 0.94910791	0.12948496
	2; 6	∓ 0.74153119	0.27970540
	3; 5	∓ 0.40584515	0.38183006
	4	0	0.41795918
8	1; 8	∓ 0.96028986	0.10122854
	2; 7	∓ 0.79666648	0.22238104
	3; 6	∓ 0.52553242	0.31370664
	4; 5	∓ 0.18343464	0.36268378

For reference (see Table 70), we give the approximate values of the abscissas of t_i and coefficients A_i in the Gaussian quadrature formula (1) for $n=1$ to 8 (see [1], [4], [6]).

The inconvenience of the Gaussian quadrature formula consists in the fact that the abscissas of the points t_i and the coefficients A_i are, generally speaking, irrational numbers. This defect is partially compensated for by its high precision for a relatively small number of ordinates.

Let us now consider using the Gaussian quadrature formula for evaluating the integral

$$\int_a^b f(x)\,dx$$

Making a change of variable,

$$x=\frac{b+a}{2}+\frac{b-a}{2}t$$

we get

$$\int_a^b f(x)\,dx=\frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)dt$$

Applying the Gaussian quadrature formula (1) to this last integral, we have

$$\int_a^b f(x)\,dx=\frac{b-a}{2}\sum_{i=1}^{n} A_i f(x_i) \tag{7}$$

where

$$x_i=\frac{b+a}{2}+\frac{b-a}{2}t_i \qquad (i=1,\ 2,\ \ldots,\ n) \tag{8}$$

and t_i are the zeros of the Legendre polynomial $P_n(t)$; that is,

$$P_n(t_i)=0$$

The remainder term of the Gaussian formula (7) with n points is expressed as follows [1], [6]:

$$R_n=\frac{(b-a)^{2n+1}(n!)^4 f^{(2n)}(\xi)}{[(2n)!]^3(2n+1)}$$

whence we obtain

$$R_2=\frac{1}{135}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi)$$

$$R_3=\frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi)$$

$$R_4=\frac{1}{3472875}\left(\frac{b-a}{2}\right)^9 f^{(8)}(\xi)$$

$$R_5 = \frac{1}{1237732650}\left(\frac{b-a}{2}\right)^{11} f^{(10)}(\xi)$$

$$R_6 = \frac{1}{648984486150}\left(\frac{b-a}{2}\right)^{13} f^{(12)}(\xi)$$

and so forth.

Example 2. Evaluate the integral

$$I = \int_0^1 \sqrt{1+2x}\,dx$$

using the Gaussian formula with three ordinates ($n=3$).

Solution. We have $a=0$ and $b=1$. By formula (8) and Table 70, the abscissas of the points will have the following values to five significant figures:

$$x_1 = \frac{1}{2} + \frac{1}{2}t_1 = 0.11270,$$

$$x_2 = \frac{1}{2} + \frac{1}{2}t_2 = 0.50000,$$

$$x_3 = \frac{1}{2} + \frac{1}{2}t_3 = 0.88730$$

The corresponding coefficients of formula (7) will, in our case, be

$$C_1 = \frac{b-a}{2}A_1 = \frac{1}{2}\cdot\frac{5}{9} = \frac{5}{18} = 0.27778$$

$$C_2 = \frac{b-a}{2}A_2 = \frac{1}{2}\cdot\frac{8}{9} = \frac{4}{9} = 0.44444$$

$$C_3 = \frac{b-a}{2}A_3 = \frac{1}{2}\cdot\frac{5}{9} = \frac{5}{18} = 0.27778$$

The subsequent computations are given in Table 71.

TABLE 71

SCHEME FOR EVALUATING AN INTEGRAL BY THE GAUSSIAN FORMULA

i	x_i	y_i	C_i	$C_i y_i$
1	0.11217	1.10698	0.27778	0.30747
2	0.50000	1.41421	0.44444	0.62853
3	0.88730	1.66571	0.27778	0.46270
Σ				1.39870

Consequently

$$I=\sum_{i=1}^{3} C_i y_i = 1.39870$$

To estimate the residual error R_3, we can take advantage of the formula

$$R_3=\frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi) \quad \text{where} \quad \xi \in (a, b)$$

Supposing

$$f(x)=\sqrt{1+2x}=(1+2x)^{\frac{1}{2}}$$

we have

$$f^{(6)}(x)=\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\left(-\frac{7}{2}\right)\left(-\frac{9}{2}\right)(1+2x)^{-\frac{11}{2}}\cdot 2^6 =$$

$$=-945(1+2x)^{-\frac{11}{2}}$$

From this,

$$\max|f^{(6)}(x)|=945 \quad \text{for} \quad 0\leqslant x\leqslant 1$$

and, hence

$$|R_3|\leqslant\frac{945}{15750}\left(\frac{1}{2}\right)^7\approx\frac{1}{2000}$$

Note that the exact value of the integral is

$$I=\sqrt{3}-\frac{1}{3}\approx 1.39872$$

16.10 SOME REMARKS ON THE ACCURACY OF QUADRATURE FORMULAS

The quadrature formulas we have considered have the following structure:

$$\int_a^b f(x)\,dx=\sum_{i=1}^{n} A_i f(x_i)+R[f] \tag{1}$$

where $x_1, x_2, \ldots, x_n$ are a given *set of points* in the interval of integration $[a, b]$, A_i are some known *constant coefficients*, and $R[f]$ is the remainder term.

The accuracy of various quadrature formulas differs for one and the same number of ordinates.

Example. Compare the accuracy of different quadrature formulas with three ordinates for the integral

$$I=\int_{-1}^{1}\sqrt{2+x}\,dx=2\sqrt{3}-\frac{2}{3}=2{,}797435\ldots$$

Solution. Using Simpson's formula, we get

$$I \approx \frac{1}{3}\left[\sqrt{2-1}+4\sqrt{2+0}+\sqrt{2+1}\right]=\frac{1}{3}\cdot 8.428905=2.809635$$

Chebyshev's formula yields the result

$$I \approx \frac{2}{3}\left[\sqrt{2-\frac{\sqrt{2}}{2}}+\sqrt{2+0}+\sqrt{2+\frac{\sqrt{2}}{2}}\right]=$$
$$=\frac{2}{3}\cdot 4.220097=2.813398$$

Finally, the Gaussian formula gives the value

$$I \approx 0.555566\left(\sqrt{2-0.774597}+\sqrt{2+0.774597}\right)+$$
$$+0.888889\sqrt{2+0}=2.797460$$

Thus, in this case the Gaussian formula is the most exact.

We confine ourselves to an examination of quadrature formulas with *equally spaced points*; these include the most common formulas: trapezoidal formula, Simpson's formula, the Newton-Cotes formula. In this case, the accuracy of the quadrature formula is in the main characterized by the *order* of the remainder term

$$R=O(h^m) \qquad (2)$$

where

$$h=\frac{b-a}{n}$$

is the spacing (n is the number of subdivisions) and m is a natural number. For example, for the trapezoidal formula (Sec. 16.3) we have

$$R[f]=-\frac{b-a}{12}h^2f''(\xi)$$

and so $m=2$; for Simpson's formula (Sec. 16.4) we have

$$R[f]=-\frac{b-a}{180}h^4f^{\mathrm{IV}}(\xi)$$

whence $m=4$. The larger the number m, the more exact the quadrature formula; in this sense, Simpson's formula is **more exact** than the trapezoidal formula. The quality of a formula is revealed when we have a sufficiently small spacing h.

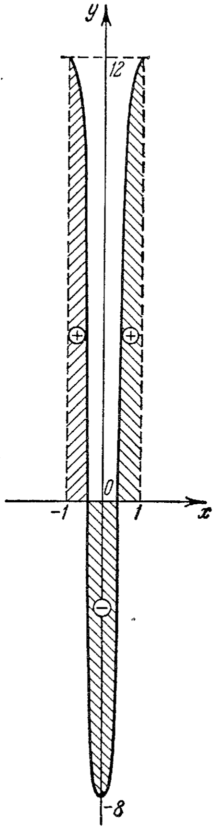


Fig. 74

From this it follows that in specific cases a crude quadrature formula cannot yield better results than a more exact one (for the same spacing). For example, for the function (Fig. 74)

$$f(x) = -8 + 45x^2 - 25x^4$$

we have

$$I = \int_{-1}^{1} f(x)\,dx = 2(-8 + 15 - 5) = 4$$

For $h = 1$ the trapezoidal formula yields the exact value

$$I_1 = \frac{1}{2} f(-1) + f(0) + \frac{1}{2} f(1) = 6 - 8 + 6 = 4$$

whereas Simpson's formula for $h = 1$ does not even ensure the sign of the integral:

$$I_2 = \frac{1}{3}[f(-1) + 4f(0) + f(1)] = \frac{1}{3}(12 - 32 + 12) = -\frac{8}{3}$$

For a fixed number of points, the accuracy of a quadrature formula depends essentially upon the location of the points. Greatly distorted values may result if the arrangement of points is not suitable. For instance, consider the function $y = f(x)$ given in Fig. 75. Choosing equally spaced points $a = x_0$, x_1, x_2, x_3, $x_4 = b$ and employing the appropriate Cotes formula for five ordinates we get

$$I = \int_a^b f(x)\,dx < 0$$

whereas it is obvious that $I > 0$.

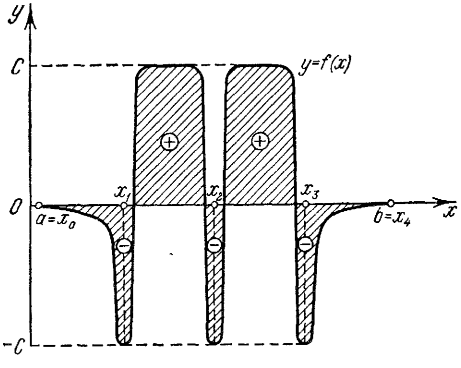


Fig. 75

It is not very difficult to construct similar examples for any quadrature formula with an arbitrary number of ordinates.

Generally, for a considerable number of zeros of the integrand $f(x)$ or for a large number of its extrema [that is, when there are many zeros of the derivative $f'(x)$], the accuracy of the quadrature formulas is greatly reduced because of the unavoidable large values of the higher derivatives. Therefore, the spacing h should be chosen so that it is much less than the distances between adjacent zeros of the function $f(x)$ and its derivative $f'(x)$. For this purpose, it is suggested to divide the basic interval of integration $[a, b]$ into subintervals $[\alpha, \beta]$, inside each of which the functions $f(x)$ and $f'(x)$ preserve sign (if this is possible), and to evaluate the integral by parts, choosing the spacing for each subinterval. In more complicated cases, one also has to take into consideration the behaviour of higher derivatives $f^{(n)}(x)$ $(n \geqslant 2)$. By way of general orientation, it is well to construct the graph of the integrand $y=f(x)$ beforehand. If the function is a highly oscillating one, it is best to employ special computational techniques. There are also general techniques which have been elaborated for increasing the accuracy of quadrature formulas [9].

To find the *total limiting error* of the quadrature formula (1), one must also take into account the *summation error* R_1. Suppose the terms $f(x_i)\,(i=1, 2, \ldots, n)$ have been computed with an absolute error not exceeding ε, and the coefficients A_i of the quadrature formula are exact positive constants. Then we can put

$$R_1 \leqslant \sum_{i=1}^{n} A_i\varepsilon = \varepsilon \sum_{i=1}^{n} A_i \qquad (3)$$

Since the quadrature formula (1) is valid for $f(x) \equiv 1$,

$$\int_a^b dx = b-a = \sum_{i=1}^{n} A_i$$

For this reason, from (3) we have

$$R_1 \leqslant (b-a)\,\varepsilon \qquad (4)$$

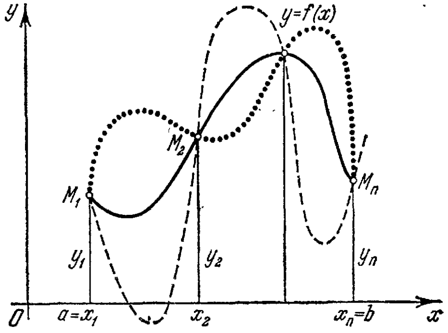


Fig. 76

Consequently, the total limiting error of a quadrature formula, without regard for the final rounding error, is

$$\bar{R} = (b-a)\,\varepsilon + |R[f]|$$

where $|R[f]|$ is the *error of method*, which may be determined as indicated above.

Note that if the integrand $y=f(x)$ is given in tabular form by the values $y_i=f(x_i)\,(i=1, 2, \ldots, n)$, then, strictly speaking, we cannot estimate the accuracy of the quadrature formula (1). This is because through a finite set of points $M_i(x_i, y_i)$ it is possible to pass an infinity of curves $y=f(x)$ (Fig. 76) bounding various areas on the given interval $[a, b]$; that is, the integral

$$I = \int_a^b f(x)\,dx$$

can a priori have an absolutely arbitrary value (see Fig. 76). In this case, quadrature formulas may be used only if we have some idea about the unused intermediate values of the integrand function and its general properties, thus enabling us to get an idea of the nature of the graph of the function.

*16.11 RICHARDSON EXTRAPOLATION

If for the quadrature formula (1) of Sec. 16.10 we know the order of the remainder term $R=R[f]$, then we can use the *method of double computation* to determine the magnitude of R. Let

$$R=O(h^m) \qquad (m \geqslant 1)$$

where

$$h=\frac{b-a}{n}$$

(n is the number of divisions), then we can approximately get

$$R=Mh^m \tag{1}$$

where M is a quantity which, for the given integrand $f(x)$, we will take to be constant on the interval of integration $[a, b]$. We choose two different spacings

$$h_1=\frac{b-a}{n_1} \quad \text{and} \quad h_2=\frac{b-a}{n_2}$$

where n_1 and n_2 $(n_2>n_1)$ are the number of subintervals in the first and second cases, respectively.

Denote by I_{n_1} and I_{n_2} the corresponding approximate values of the integral I. From (1) we have

$$R_{n_1}=I-I_{n_1}=M\left(\frac{b-a}{n_1}\right)^m \tag{2}$$

and

$$R_{n_2}=I-I_{n_2}=M\left(\frac{b-a}{n_2}\right)^m \tag{2'}$$

where R_{n_1} and R_{n_2} are the appropriate remainder terms. Then

$$I_{n_2}-I_{n_1}=M(b-a)^m\left(\frac{1}{n_1^m}-\frac{1}{n_2^m}\right)$$

and, consequently,

$$M=\frac{(n_1n_2)^m}{(b-a)^m}\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m}$$

On the basis of (1) we get an expression for the remainder term:

$$R=\left(\frac{n_1n_2}{n}\right)^m\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m}$$

in particular, for $h = h_2$; that is, when $n = n_2$ we have

$$R_{n_2} = \frac{n_1^m}{n_2^m - n_1^m}(I_{n_2} - I_{n_1}) \tag{3}$$

Using correction (3) and by virtue of (2′), we get the following improved value for the integral I:

$$I_{n_1, n_2} = I_{n_2} + \frac{n_1^m}{n_2^m - n_1^m}(I_{n_2} - I_{n_1}) \tag{4}$$

This technique is called *Richardson extrapolation* [10]. Introducing the notation

$$\frac{n_2}{n_1} = \alpha \quad (\alpha > 1)$$

we have

$$I_{n_1, n_2} = I_{n_2} + \beta (I_{n_2} - I_{n_1}) \tag{5}$$

where

$$\beta = \frac{1}{\alpha^m - 1} \tag{6}$$

The coefficients β are tabulated for various values of α and m. Note that for the trapezoidal formula $m = 2$ and for Simpson's formula, $m = 4$. A special case of formula (5) was given in Sec. 16.7.

We will show that if $I_{n_1} \neq I_{n_2}$, then I_{n_1, n_2} always lies outside the interval $[I_{n_1}, I_{n_2}]$.

Indeed, if

$$I_{n_2} > I_{n_1}$$

then from formula (5) we have

$$I_{n_1, n_2} > I_{n_2} = \max \{I_{n_1}, I_{n_2}\}$$

But if

$$I_{n_2} < I_{n_1}$$

then from the same formula (5) we obtain

$$I_{n_1, n_2} = I_{n_2} - \beta (I_{n_1} - I_{n_2}) < I_{n_2} = \min \{I_{n_1}, I_{n_2}\}$$

Thus

$$I_{n_1, n_2} \notin [I_{n_1}, I_{n_2}]$$

That is, I_{n_1, n_2} is obtained from I_{n_1} and I_{n_2} by means of an **extrapolation** operation, whence the name of the method.

If $I_{n_1} = I_{n_2}$, then obviously

$$I_{n_1, n_2} = I_{n_1} = I_{n_2}$$

It may be shown that for a sufficiently smooth integrand function $f(x)$ the order of the remainder term for I_{n_1, n_2} is at least equal to $m+1$.

Note. Tables 72a and 72b give examples of Richardson extrapolation.

TABLE 72a

EXTRAPOLATION FOR THE CASE OF THE TRAPEZOIDAL FORMULA

No.		I_2	I_4	$I_{2, 4}$	I
1	$I=\int_0^{\pi} \sin x\, dx$	1.571	1.896	2.004	2.000
2	$I=\int_0^2 e^{-x^2}\, dx$	0.877	0.881	0.8823	0.8821
3	$I=\int_3^7 x^2 \ln x\, dx$	185.7090	179.5385	177.4819	177.4836
4	$I=\int_0^4 \frac{dx}{\sqrt{5-x^2}}$	0.9695	0.9389	0.9286	0.9267

No.	$e_1=I-I_2$	$e_2=I-I_4$	$e_{1,2}=I-I_{2,4}$
1	0.429	0.104	—0.004
2	0.0051	0.0011	—0.0002
3	—8.2254	—2.0549	0.0017
4	—0.0428	—0.0122	—0.0019

From these two tables it will be seen that, as a rule, the extrapolation increases the accuracy of computations for functions without singularities.

It is also possible to derive more exact extrapolation formulas using the values I_{n_1}, I_{n_2} and I_{n_3} of the desired integral that correspond to the three distinct spacings

$$h_s=\frac{b-a}{n_s} \quad (s=1,\ 2,\ 3)$$

and taking into account the first two terms of the expansion of the remainder term of the quadrature formula [10].

TABLE 72B

EXTRAPOLATION FOR THE CASE OF SIMPSON'S FORMULA

No.		I_2	I_4	$I_{2,4}$	I
1	$I=\int_0^{\pi}\sin x\,dx$	2.094	2.004	2.010	2.000
2	$I=\int_0^1\frac{dx}{1+x^2}$	0.7833	0.7853	0.7855	0.7854
3	$I=\int_3^7 x^2\ln x\,dx$	177.454	177.481	177.483	177.4836
4	$I=\int_0^4\frac{dx}{(25-x^2)^{3/2}}$	0.0577	0.0541	0.0538	0.0533

No.	$e_1=I-I_1$	$e_2=I-I_4$	$e_{1,2}=I-I_{2,4}$
1	−0.094	−0.004	−0.010
2	0.0021	0.0001	−0.0001
3	0.0296	0.0026	0.0006
4	−0.0044	−0.0008	−0.0005

*16.12 BERNOULLI NUMBERS

Consider the function

$$f(x)=\frac{x}{e^x-1} \tag{1}$$

Taking advantage of the familiar expansion

$$e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots$$

we can write

$$f(x)=\frac{x}{\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots}=\frac{1}{1+\frac{x}{2!}+\frac{x^2}{3!}+\dots} \tag{2}$$

From this it is clear that the function $f(x)$ can be expanded in a power series about $x=0$; for the sake of convenience in subsequent computations, we represent this series as

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n \tag{3}$$

where $B_0 = f(0) = 1$. In order to determine the other coefficients B_n $(n = 1, 2, \ldots)$ of the expansion, which are called *Bernoulli numbers*, we make use of the identity obtained from (2):

$$\sum_{n=0}^{\infty} \frac{x^n}{(n+1)!} \cdot \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n \equiv 1$$

Multiplying together the power series and equating to zero the coefficients of positive powers of the variable x, we obtain an infinite system of linear equations:

$$\frac{B_n}{n!} \cdot \frac{1}{1!} + \frac{B_{n-1}}{(n-1)!} \cdot \frac{1}{2!} + \ldots + \frac{B_0}{0!} \frac{1}{(n+1)!} = 0 \qquad (n = 1, 2, 3 \ldots)$$

or, multiplying by $(n+1)!$ and noting that

$$\frac{(n+1)!}{(n-k)!\,(k+1)} = C_{n+1}^{n-k} \qquad (k = 0, 1, \ldots, n+1)$$

we get

$$C_{n+1}^1 B_n + C_{n+1}^2 B_{n-1} + \ldots + C_{n+1}^n B_1 + 1 = 0 \tag{4}$$

If we agree to set

$$B_k = B^k \tag{5}$$

then formula (4) may be compactly written in the following *symbolic* form:

$$(B+1)^{n+1} - B^{n+1} + 0$$

or, replacing $n+1$ by n,

$$(B+1)^n - B^n = 0 \tag{6}$$

Putting $n = 2, 3, 4 \ldots$ in formula (6), we obtain an infinite system of equations:

$$\left.\begin{aligned} 2B_1 + 1 &= 0, \\ 3B_2 + 3B_1 + 1 &= 0, \\ 4B_3 + 6B_2 + 4B_1 + 1 &= 0, \\ 5B_4 + 10B_3 + 10B_2 + 5B_1 + 1 &= 0, \\ \cdots\cdots\cdots\cdots\cdots & \end{aligned}\right\} \tag{7}$$

Whence, we successively find

$$B_1=-\frac{1}{2},\quad B_2=\frac{1}{6},\quad B_3=0,\quad B_4=-\frac{1}{30},\quad B_5=0,$$
$$B_6=\frac{1}{42},\quad B_7=0,\quad B_8=-\frac{1}{30},\quad B_9=0,\quad B_{10}=\frac{5}{66},$$
$$B_{11}=0,\quad B_{12}=-\frac{691}{2730},\quad B_{13}=0,\quad B_{14}=\frac{7}{6},\quad B_{15}=0,$$
$$B_{16}=-\frac{3617}{510},\quad B_{17}=0,\quad B_{18}=\frac{43\,867}{798},\quad B_{19}=0,$$
$$B_{20}=-\frac{174\,611}{330}$$

and so on.

Thus, the Bernoulli numbers may be determined step by step from the symbolic formula (6); note that after the binomial expansion the powers of the B numbers must be replaced by Bernoulli numbers with the appropriate indices.

Function (1) is called the *generating function* of Bernoulli numbers. Taking advantage of the notation of (5), we can write expansion (3) symbolically as follows:

$$\frac{x}{e^x-1}=e^{Bx}$$

It is obvious from the structure of system (7) that all the Bernoulli numbers are rational. Besides, it turns out that the Bernoulli numbers of odd index, with the exception of B_1, are equal to zero. We prove this property in the general form. Noting that

$$B_0=1 \quad \text{and} \quad B_1=-1/2$$

we have

$$\varphi(x)=\frac{x}{e^x-1}-B_1x=\frac{x}{e^x-1}+\frac{x}{2}=1+\sum_{n=2}^{\infty}\frac{B_n}{n!}x^n \tag{8}$$

Clearly

$$\varphi(x)=\frac{x(e^x+1)}{2(e^x-1)}=\frac{x}{2}\cdot\frac{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}{e^{\frac{x}{2}}-e^{-\frac{x}{2}}}=\frac{x}{2}\ \coth\ \frac{x}{2}$$

is an even function, and so its expansion (8) contains only even powers of the variable x and, hence,

$$B_n=0 \quad \text{for} \quad n=3,\ 5,\ 7,\ \ldots$$

Bernoulli numbers find applications in a diversity of problems. For instance, they are used in the important *Euler-Maclaurin summation formula* which we will now derive.

*16.13 EULER-MACLAURIN FORMULA

Let $y=f(x)$ be a function defined in the domain $x \geqslant x_0$. We consider the *finite-difference* operator

$$\Delta f(x)=f(x+h)-f(x)$$

where h is a fixed positive quantity. The *inverse operator* $\frac{1}{\Delta}$ of the function $f(x)$ is naturally taken to mean the function $F(x)$ which satisfies the finite-difference equation

$$\Delta F(x)=f(x) \tag{1}$$

Thus, from (1) we have

$$F(x)=\frac{1}{\Delta} f(x) \tag{2}$$

If $f(x)$ is regarded on a set of equally spaced points

$$x_0, \ x_1, \ x_2, \ \ldots$$

where $\Delta x_i = x_{i+1}-x_i=h$ $(i=0, 1, 2, \ldots)$, then the inverse operator $F(x_i)=\frac{1}{\Delta} f(x_i)$ can easily be constructed. Indeed, we form the finite sum

$$S(x_i)=\sum_{j=0}^{i-1} f(x_j) \qquad (i=1, 2, \ldots)$$

and agree that $S(x_0)=0$. We clearly get

$$\Delta S(x_i)=S(x_{i+1})-S(x_i)=f(x_i) \tag{3}$$

On the other hand, by (1) we have

$$\Delta F(x_i)=f(x_i) \tag{4}$$

Subtracting (4) from (3), we get

$$\Delta [F(x_i)-S(x_i)]=0$$

for $i=0, 1, 2, \ldots$. Here, the difference $F(x_i)-S(x_i)$ does not depend on the subscript i and we can put

$$F(x_i)-S(x_i)=F(x_0)-S(x_0)=F \ x_0)$$

whence

$$F(x_i)=F(x_0)+S(x_i)$$

where $F(x_0)$ is an arbitrary constant. Thus

$$\frac{1}{\Delta} f(x_i)=F(x_0)+S(x_i) \tag{5}$$

That is to say, *the inverse operator for a finite difference is the operator of a finite summation.*

Let us now introduce the *differentiation operator*

$$Df(x)=\frac{df(x)}{dx}$$

The *inverse operator* $\frac{1}{D}$ is to be understood in the sense of the operation of integration:

$$\frac{1}{D}f(x)=\int_{x_0}^{x} f(x)\,dx$$

Using the Taylor series, we find

$$\Delta f(x)=\sum_{k=1}^{\infty}\frac{h^k}{k!}D^kf(x)=\left\{\sum_{k=1}^{\infty}\frac{h^kD^k}{k!}\right\}f(x)=(e^{hD}-1)f(x)$$

Hence

$$\Delta=(e^{hD}-1)$$

Then, for the inverse operator $\frac{1}{\Delta}$, we get the following expression:

$$\frac{1}{\Delta}=\frac{1}{e^{hD}-1}$$

Multiplying both members of this equation by hD, we get

$$hD\frac{1}{\Delta}=\frac{hD}{e^{hD}-1}$$

In the right-hand member we have the generating function of Bernoulli numbers and so

$$hD\frac{1}{\Delta}=\sum_{k=0}^{\infty}\frac{B_k}{k!}h^kD$$

or, expanded,

$$\frac{d}{dx}\left[\frac{1}{\Delta}f(x)\right]=\sum_{k=0}^{\infty}\frac{B_k}{k!}h^{k-1}D^kf(x) \qquad (6)$$

Integrating (6) from $x=x_0$ to $x=x_n$ and using formula (5), we obtain

$$\frac{1}{\Delta}f(x_n)-\frac{1}{\Delta}f(x_0)=$$

$$=\frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx+\sum_{k=1}^{\infty}\frac{B_k}{k!}h^{k-1}\left[f^{(k-1)}(x_n)-f^{(k-1)}(x_0)\right]$$

or

$$F(x_0)+\sum_{i=0}^{n-1} f(x_i)-F(x_0)=\sum_{j=0}^{n-1} f(x_j)=$$

$$=\frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx+\sum_{k=1}^{\infty}\frac{B_k}{k!}h^{k-1}\left[f^{(k-1)}(x_n)-f^{(k-1)}(x_0)\right]$$

Taking into account that

$$B_1=-\frac{1}{2} \quad \text{and} \quad B_{2k+1}=0 \quad \text{for} \quad k=1, 2, \ldots$$

we get the *Euler-Maclaurin formula* (also called the Euler-Maclaurin summation formula)

$$\int_{x_0}^{x_n} f(x)\,dx=h\left[\frac{1}{2}f(x_0)+f(x_1)+f(x_2)+\ldots+f(x_{n-1})+\frac{1}{2}f(x_n)\right]-$$

$$-\sum_{k=1}^{m}\frac{B_{2k}}{(2k)!}h^{2k}\left[f^{(2k-1)}(x_n)-f^{(2k-1)}(x_0)\right]+R_{2m} \qquad (7)$$

where R_{2m} is the *remainder term*. The notation (7) in the form of an infinite series is not always legitimate since the series may diverge. Substituting the values of the Bernoulli numbers, we get

$$\int_{x_0}^{x_n} y\,dx=h\left(\frac{1}{2}y_0+y_1+y_2+\ldots+y_{n-1}+\frac{1}{2}y_n\right)-\frac{h^2}{12}(y'_n-y'_0)+$$

$$+\frac{h^4}{720}(y'''_n-y'''_0)-\frac{h^6}{30240}(y^{V}_n-y^{V}_0)+\ldots$$

$$\ldots-\frac{B_{2m}}{(2m)!}h^{2m}\left[f^{(2m-1)}(x_n)-f^{(2m-1)}(x_0)\right]+R_{2m} \quad (8)$$

The remainder term in the Euler-Maclaurin formula is of the form [6]

$$R_{2m}=-nh^{2m+3}\frac{B_{2m+2}}{(2m+2)!}f^{(2m+2)}(\xi)$$

where $\xi\in(x_0, x_n)$.

The Euler-Maclaurin formula (8) is used for approximate evaluation of definite integrals and also for approximate summation of the values of functions for equally spaced values of the argument.

Indeed, from (8) we have

$$\sum_{i=0}^{n} f(x_i) = \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx + \frac{f(x_0)+f(x_n)}{2} + \\ + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} h^{2k-1}\left[f^{(2k-1)}(x_n) - f^{(2k-1)}(x_0)\right] + \\ + nh^{2m+2}\frac{B_{2m+2}}{(2m+2)!} f^{(2m+2)}(\xi) \qquad (9)$$

Example 1. Using the Euler-Maclaurin formula, calculate the approximate value of the definite integral

$$I = \int_{0.2}^{1} (\sin x - \ln x + e^x)\,dx$$

Solution. Divide the interval [0.2, 1] into, say, eight subintervals taking $h = 0.1$ and setting

$$x_i = 0.2 + i\cdot 0.1 \quad (i = 0, 1, \ldots, 8)$$

The results of the computations of the values of the function $f(x) = \sin x - \ln x + e^x$ are listed in Table 73.

TABLE 73

VALUES OF THE FUNCTION $f(x) = \sin x - \ln x + e^x$

x	0.2	0.3	0.4	0.5	0.6
$f(x)$	3.02951	2.84936	2.79754	2.82130	2.89759
x	0.7	0.8	0.9	1.0	
$f(x)$	3.01435	3.16605	3.34830	3.55975	

whence

$$\frac{1}{2}f(x_0) + f(x_1) + \ldots + f(x_7) + \frac{1}{2}f(x_8) = 24.1894$$

Confining ourselves to the fifth derivative, we have

$$f'(x) = \cos x - \frac{1}{x} + e^x,$$

$$f'''(x) = -\cos x - \frac{2}{x^3} + e^x,$$

$$f^{\mathrm{V}}(x) = \cos x - \frac{24}{x^5} + e^x$$

Hence

$$f'(0.2) = -2.7985, \qquad f'(1) = 2.2586,$$
$$f'''(0.2) = -249.7587 \qquad f'''(1) = 0.1780$$
$$f^{\mathrm{V}}(0.2) = -74997.7985 \qquad f^{\mathrm{V}}(1) = -20.7415$$

Substituting the values found into (8), we obtain

$$I = 24.1894 \cdot 0.1 - \frac{(0.1)^2}{12} \cdot (2.2586 + 2.7985) + \frac{(0.1)^4}{720} \cdot (0.1780 + 249.7587) - \frac{(0.1)^6}{30240} \cdot (-20.7415 + 74997.7985) = 2.41894 - 0.00421 + 0.00004 = 2.41477$$

Direct integration yields

$$I = [-\cos x - x(\ln x - 1) + e^x] \Big|_{0.2}^{1} \approx 2.4148$$

Example 2. Find the sum of

$$\frac{1}{51^2} + \frac{1}{53^2} + \frac{1}{55^2} + \ldots + \frac{1}{99^2}$$

Solution. In our case

$$f(x) = \frac{1}{x^2}, \quad h = 2, \quad x_0 = 51, \quad x_n = 99$$

We find derivatives of odd order of the function $f(x)$:

$$f'(x) = -\frac{2}{x^3},$$

$$f'''(x) = -\frac{24}{x^5},$$

$$f^{\mathrm{V}}(x) = -\frac{720}{x^7},$$

$$f^{\mathrm{VII}}(x) = -\frac{40320}{x^9}, \text{ etc.}$$

Substituting into formula (9) and restricting ourselves to the seventh derivative, we obtain

$$\sum_{x=51}^{x=99} \frac{1}{x^2} = \frac{1}{2} \int_{51}^{99} \frac{dx}{x^2} + \frac{1}{2}\left(\frac{1}{51^2} + \frac{1}{99^2}\right) + \frac{1}{3}\left(\frac{1}{51^3} - \frac{1}{99^3}\right) - \frac{4}{15}\left(\frac{1}{51^5} - \frac{1}{99^5}\right) + \frac{16}{21}\left(\frac{1}{51^7} - \frac{1}{99^7}\right) - \frac{64}{15}\left(\frac{1}{51^9} - \frac{1}{99^9}\right) =$$
$$= 0.004753416 + 0.000243490 + 0.000002169 - 0.000000001 = 0.004999074$$

By (9), where we have to put $h=2$, $n=24$, $m=4$, the error in the result is

$$R=24\cdot 2^{10}\cdot\frac{B_{10}}{8!}\cdot f^{(10)}(\xi)<24\cdot 2^{10}\cdot\frac{5}{66}\cdot\frac{1}{8!}\cdot\frac{11!}{50^{12}}<\frac{2}{25^{10}}\approx 10^{-14}$$

16.14 APPROXIMATION OF IMPROPER INTEGRALS

An integral

$$\int_a^b f(x)\,dx \tag{1}$$

is called *proper* if (1) the interval of integration $[a, b]$ is finite, (2) the integrand $f(x)$ is continuous on $[a, b]$, otherwise the integral (1) is termed *improper*.

Let us first consider the approximate computation of the improper integral

$$\int_a^\infty f(x)\,dx \tag{2}$$

with an *infinite interval of integration* where the function $f(x)$ is continuous over $a\leqslant x<\infty$.

The integral (2) is *convergent* if there is a finite limit

$$\lim_{b\to\infty}\int_a^b f(x)\,dx \tag{3}$$

and, by definition, we assume

$$\int_a^\infty f(x)\,dx=\lim_{b\to\infty}\int_a^b f(x)\,dx \tag{4}$$

If the limit (3) does not exist, then the integral (2) is *divergent*; such an integral is considered to be meaningless. Therefore, before attempting to evaluate an improper integral one must make sure, using familiar convergence tests [10], that the integral converges.

To evaluate the convergent improper integral (2) to a given accuracy ε, we represent it in the form

$$\int_a^\infty f(x)\,dx=\int_a^b f(x)\,dx+\int_b^\infty f(x)\,dx \tag{5}$$

Since the integral converges, the number b may be chosen so large that the inequality

$$\left|\int_{b}^{\infty} f(x)\,dx\right| < \frac{\varepsilon}{2} \tag{6}$$

is valid.

The proper integral

$$\int_{a}^{b} f(x)\,dx$$

may be computed from one of the quadrature formulas. Let S be an approximate value of this integral to an accuracy of $\frac{\varepsilon}{2}$; thus

$$\left|\int_{a}^{b} f(x)\,dx - S\right| < \frac{\varepsilon}{2} \tag{7}$$

From formulas (5), (6) and (7) we have

$$\left|\int_{a}^{\infty} f(x)\,dx - S\right| < \varepsilon$$

and so the problem can be solved.

Suppose now that the interval of integration $[a, b]$ is finite and the integrand $f(x)$ has a finite number of discontinuities on $[a, b]$. Since under our assumptions the interval of integration may be partitioned into subintervals with a single discontinuity point of the integrand function, it suffices to examine one case: when there is a single discontinuity point c of the function $f(x)$ on $[a, b]$, this point being of the second kind.[1]

If c is an interior point of $[a, b]$, then by definition we put

$$\int_{a}^{b} f(x)\,dx = \lim_{\substack{\delta_1 \to +0 \\ \delta_2 \to +0}} \left\{ \int_{a}^{c-\delta_1} f(x)\,dx + \int_{c+\delta_2}^{b} f(x)\,dx \right\} \tag{8}$$

---

[1] If c is a point of discontinuity of the first kind, that is, there exist finite one-sided limits

$$f(c-0) = \lim_{x \to c,\ x<c} f(x) \quad \text{and} \quad f(c+0) = \lim_{x \to c,\ x>c} f(x),$$

then we can put

$$\int_{a}^{b} f(x)\,dx = \int_{a}^{c} f_1(x)\,dx + \int_{c}^{b} f_2(x)\,dx$$

where

$$f_1(x) = \begin{cases} f(x) & \text{if } a \leqslant x < c \\ f(c-0) & \text{if } x = c \end{cases} \quad \text{and} \quad f_2(x) = \begin{cases} f(c+0) & \text{if } x = c \\ f(x) & \text{if } c < x \leqslant b \end{cases}$$

The functions $f_1(x)$ and $f_2(x)$ are continuous on the intervals $[a, c]$ and $[c, b]$ respectively. Thus, our integral reduces to the sum of two proper integrals.

and if the limit exists the integral is *convergent*, otherwise it is *divergent*.

In similar fashion we define the convergence of the improper integral (8) if the discontinuity point c of the integrand $f(x)$ coincides with one of the endpoints of the interval of integration $[a, b]$.

In order to approximate, to a given accuracy ε, the convergent improper integral (8), where the point of discontinuity $c \in (a, b)$, one chooses positive numbers δ_1 and δ_2 so small that the inequality

$$\left| \int_{c-\delta_1}^{c+\delta_2} f(x)\,dx \right| < \frac{\varepsilon}{2}$$

holds true. Then, using familiar quadrature formulas, one approximately calculates the proper integrals

$$\int_{a}^{c-\delta_1} f(x)\,dx \quad \text{and} \quad \int_{c+\delta_2}^{b} f(x)\,dx \tag{9}$$

It is clear that if S_1 and S_2 are approximate values of the integrals (9) to within $\frac{\varepsilon}{4}$, then

$$\int_{a}^{b} f(x)\,dx \approx S_1 + S_2$$

to an accuracy of ε. If the discontinuity point c of the integrand $f(x)$ is an endpoint of the interval of integration $[a, b]$, then the computational procedure is modified accordingly.

16.15 THE METHOD OF KANTOROVICH FOR ISOLATING SINGULARITIES

A useful device in calculating the approximate value of an integral of a discontinuous function is the *method of L. V. Kantorovich for isolating singularities* [1], [6]. [10]. The underlying principle of this method is that we take out of the integrand $f(x)$ a certain function $g(x)$ having the same singularities as $f(x)$, which is integrable in elementary terms on the given interval $[a, b]$ and is such that the difference $f(x)-g(x)$ is sufficiently smooth on $[a, b]$. For example,

$$f(x)-g(x) \in C^{(m)}[a, b], \quad \text{where} \quad m \geqslant 1$$

We then have

$$\int_a^b f(x)\,dx = \int_a^b g(x)\,dx + \int_a^b [f(x)-g(x)]\,dx$$

where the first integral is taken directly, and the second integral is readily evaluated by means of standard formulas.

Let us consider the use of this method for computing integrals of the form

$$\int_a^b \frac{\varphi(x)}{(x-x_0)^{\alpha}}\,dx \tag{1}$$

where $x_0 \in [a, b]$, $0<\alpha<1$ and $\varphi(x)$ is continuous on $[a, b]$.

Let $\varphi(x) \in C^{(m+1)}[a, b]$, that is, let $\varphi(x)$ have continuous derivatives on $[a, b]$ up to the order $(m+1)$ inclusive.

Using Taylor's formula, we have

$$\varphi(x) = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}(x-x_0)^k + \psi(x) \tag{2}$$

where

$$\psi(x) = \varphi(x) - \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}(x-x_0)^k = \frac{\varphi^{(m+1)}(\xi)}{(m+1)!}(x-x_0)^{m+1} \tag{3}$$

$$[\xi \in (a, b)]$$

From this, we get, for integral (1),

$$\int_a^b \frac{\varphi(x)\,dx}{(x-x_0)^{\alpha}} = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}\int_a^b (x-x_0)^{k-\alpha}\,dx + \int_a^b \frac{\psi(x)\,dx}{(x-x_0)^{\alpha}} =$$

$$= \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!\,(k+1-\alpha)}\left[(b-x_0)^{k+1-\alpha} - (a-x_0)^{k+1-\alpha}\right] + I \tag{4}$$

where

$$I = \int_a^b \frac{\psi(x)\,dx}{(x-x_0)^{\alpha}} \tag{5}$$

From formula (3) it follows that

$$\frac{\psi(x)}{(x-x_0)^{\alpha}} \in C^{(m)}[a, b]$$

(at least!); hence, integral (5) is a proper integral and can be computed to any degree of accuracy by using the appropriate quadrature formula.

The Kantorovich method is also applicable to improper integrals whose integrands have several points of discontinuity of the type considered. In that case, to evaluate the integral it suffices to partition the interval of integration into parts containing only one singular point of the integrand function, and then take advantage of the additive property of integrals.

Example 1. Calculate the approximate value of the improper integral [11]

$$I=\int_0^{\frac{1}{2}}\frac{dx}{\sqrt{x(1-x)}}$$

Solution. The integrand

$$f(x)=x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}}$$

has one singularity $x=0$ on the interval $\left[0, \frac{1}{2}\right]$. Expand the function

$$\varphi(x)=(1-x)^{-\frac{1}{2}}$$

in a Taylor series in powers of x up to x^4. Using the binomial theorem, we have

$$\varphi(x)=1+\frac{1}{2}x+\frac{3}{8}x^2+\frac{5}{16}x^3+\frac{35}{128}x^4$$

whence

$$I=\int_0^{\frac{1}{2}}x^{-\frac{1}{2}}\,dx+\frac{1}{2}\int_0^{\frac{1}{2}}x^{\frac{1}{2}}\,dx+$$

$$+\frac{3}{8}\int_0^{\frac{1}{2}}x^{\frac{3}{2}}\,dx+\frac{5}{16}\int_0^{\frac{1}{2}}x^{\frac{5}{2}}\,dx+\frac{35}{128}\int_0^{\frac{1}{2}}x^{\frac{7}{2}}\,dx+I_1=\frac{715801}{645120}\sqrt{2}+I_1=$$

$$=1{,}5691585+I_1, \quad (6)$$

where

$$I_1=\int_0^{\frac{1}{2}}\frac{\psi(x)}{\sqrt{x}}\,dx \quad (7)$$

and

$$\psi(x)=\frac{1}{\sqrt{1-x}}-\left(1+\frac{1}{2}x+\frac{3}{8}x^2+\frac{5}{16}x^3+\frac{35}{128}x^4\right); \quad \psi(0)=0$$

We compute the proper integral (7) by Simpson's formula taking $n=10$ and spacing $h=\frac{1}{20}=0.05$. The results of the computations to six decimal places are listed in Table 74.

TABLE 74

EVALUATING INTEGRAL (7) BY SIMPSON'S FORMULA

i	x_i	y_{2j-1}	y_{2j}
0	0		$\boxed{0.000000}$
1	0.05		
2	0.10	0.000000	0.000009
3	0.15		
4	0.20	0.000056	0.000216
5	0.25		
6	0.30	0.000624	0.001508
7	0.35		
8	0.40	0.003225	0.006316
9	0.45		
10	0.50	0.011588	$\boxed{0.020239}$
Σ		0.015493	0.008049

From this, we have

$$I_1=\frac{1}{20\cdot 3}(0.020239+4\cdot 0.015493+2\cdot 0.008049)=\frac{1}{60}\times$$
$$\times 0.098309=0.0016385$$

Hence, by (6), we have

$$I=+\begin{matrix}1.5691585\\0.0016385\end{matrix}\Bigg\}=1.5707970$$

Note that the integral I is expressible in elementary terms and its exact value is

$$I=\frac{\pi}{2}=1.5707963\ldots$$

Note. In some cases, an improper integral can be transformed into a proper integral by a change of variable or by integration by parts

Example 2. Transform

$$I=\int_{1}^{\infty}\frac{dx}{(1+x)\sqrt{x}} \tag{8}$$

into a proper integral.

Solution. Putting $x=\frac{1}{z}$ in (8), we obtain an integral with finite limits:

$$I=\int_{0}^{1}\frac{dz}{(z+1)\sqrt{z}}=\int_{0}^{1}\frac{dx}{(1+x)\sqrt{x}} \tag{9}$$

but with a singularity at $x=0$.

Integrating by parts, we have

$$I=\int_{0}^{1}\frac{1}{1+x}d(2\sqrt{x})=\frac{2\sqrt{x}}{1+x}\Big|_{0}^{1}+\int_{0}^{1}2\sqrt{x}\frac{dx}{(1+x)^2}=$$

$$=1+2\int_{0}^{1}\frac{\sqrt{x}}{(1+x)^2}dx$$

The remaining integral is a proper integral, and quadrature formulas can be applied to it without any difficulties. Other techniques are also used for the transformation of improper integrals [6].

16.16 GRAPHICAL INTEGRATION

The problem of graphical integration consists in the following: from the given graph of a continuous function $y=f(x)$ it is required to construct the graph of its antiderivative:

$$F(x)=\int_{a}^{x}f(x)\,dx$$

In other words, we have to construct a curve $y=F(x)$ whose ordinate at each point x is numerically equal to the area of a curvilinear trapezoid with base $[a, x]$ bounded by the given curve $y=f(x)$.

For an approximate construction of the graph of the antiderivative $y=F(x)$, we partition the area of the corresponding curvilinear trapezoid bounded by the curve $y=f(x)$ into narrow vertical strips using the ordinates drawn at the points $x_0, x_1, \ldots$ $(a=x_0<x_1<x_2<\ldots)$ (Fig. 77). Using the mean-value theorem, we replace each of these strips with an equal (as far as possible) rectangle having the same base and with altitude equal to $f(\xi_i)$ where $\xi_i (i = 1, 2, \dots)$ is an intermediate point of the ith interval $[x_{i-1}, x_i]$; thus, we put

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(\xi_i)(x_i - x_{i-1})$$

where

$$x_{i-1} \leqslant \xi_i \leqslant x_i \qquad (i = 1, 2, \dots)$$

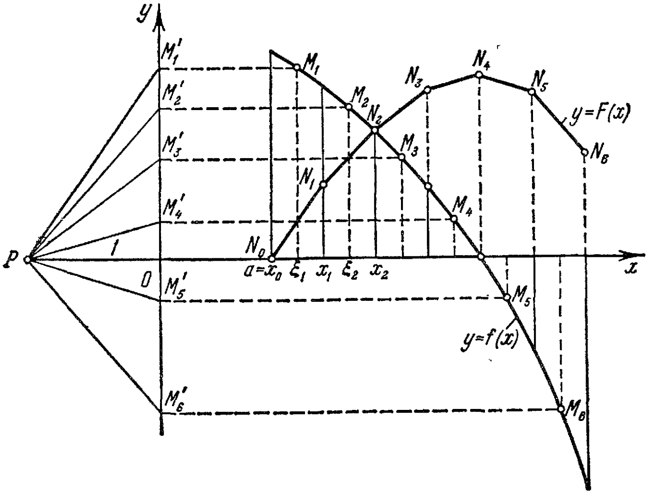


Fig. 77

The values of the antiderivative

$$F(x) = \int_{x_0}^{x} f(x)\,dx$$

may be computed at the points x_i by the method of accumulation: $F(x_0) = 0$,

$$F(x_i) = \int_{x_0}^{x_i} f(x)\,dx = \int_{x_0}^{x_{i-1}} f(x)\,dx + \int_{x_{i-1}}^{x_i} f(x)\,dx =$$
$$= F(x_{i-1}) + f(\xi_i)(x_i - x_{i-1}) \quad (i = 1, 2, \dots) \qquad (1)$$

Let $M_1(\xi_1, f(\xi_1))$, $M_2(\xi_2, f(\xi_2))$, ... be the corresponding points of the curve $y = f(x)$. Projecting them on the y-axis, we obtain the points M_1', M_2'', ... (Fig. 77).

Now choose pole P with the distance $OP = 1$ and draw rays PM_1', PM_2', ... The desired curve $y = F(x)$ may be approximately replaced by the polygonal line $N_0N_1N_2N_3$ ... with vertices

$N_0(x_0, 0)$, $N_1(x_1, F(x_1))$, $N_2(x_2, F(x_2))$, ... The successive segments of this polygonal line will be parallel to the corresponding rays; namely, $N_0N_1 \| PM_1'$, $N_1N_2 \| PM_2'$, $N_2N_3 \| PM_3'$, ... Indeed, on the basis of formula (1), the slope of the segment $N_{i-1}N_i$ is

$$k = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}} = f(\xi_i)$$

Now, by virtue of the construction, the slope of ray OM_i' is

$$k_i' = \frac{f(\xi_i)}{1} = f(\xi_i)$$

Hence

$$N_{i-1}N_i \| OM_i' \qquad (i = 1, 2, \ldots)$$

Thus, the actual construction of the graph of the function $y = F(x)$ may be carried out as follows: from point $N_0(x_0, 0)$ draw the straight line N_0N_1 parallel to the ray OM_1' to intersection with the vertical line $x = x_1$ at the point N_1; from N_1 draw the straight line N_1N_2 parallel to the ray OM_2' to intersection, at N_2, with the vertical $x = x_2$ and so on.

Note that when applying this method of graphical integration the points $x_i (i = 0, 1, \ldots)$ need not be taken equally spaced. To increase the accuracy of the construction, it is advisable to include in the set of points x_i the characteristic points of the graph of the function to be integrated (zeros, extrema, points of inflection).

Generally speaking, graphical integration has a low degree of accuracy and so is useful when the aim is to obtain a rough picture of the integral of the function or when the integrand is specified graphically and its analytical expression is not known.

*16.17 ON CUBATURE FORMULAS

Cubature formulas are designed for numerical evaluation of double integrals [1].

Suppose a function $z = f(x, y)$ is defined and continuous in some bounded domain σ (Fig. 78). In this domain σ we choose a set of points (lattice points) $M_i(x_i, y_i)$ $(i = 1, 2, \ldots, N)$. To compute the double integral $\iint\limits_{(\sigma)} f(x, y)\,dx\,dy$ we approximately put

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{N} A_i f(x_i, y_i) \qquad (1)$$

In order to find the coefficients A_i we will require that the cubature formula (1) hold true for all polynomials

$$P_n(x, y) = \sum_{k+l \leqslant n} C_{kl} x^k y^l \qquad (2)$$

whose degree does not exceed a specified number n. For this it is necessary and sufficient that formula (1) be exact for the product of powers

$$x^k y^l$$
$$(k,\ l=0,\ 1,\ 2,\ \ldots,\ n;\quad k+l \leqslant n)$$

Putting $f(x,\ y)=x^k y^l$ in (1), we have

$$I_{kl}=\iint\limits_{(\sigma)} x^k y^l\,dx\,dy=\sum_{i=1}^{N} A_i x_i^k y_i^l$$
$$(k,\ l=0,\ 1,\ 2,\ \ldots,\ n,\ k+l \leqslant n) \tag{3}$$

Thus, generally speaking, the coefficients A_i of (1) can be determined from the system of linear equations (3).

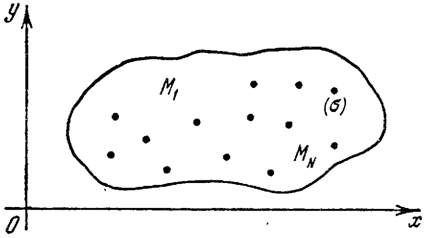


Fig. 78

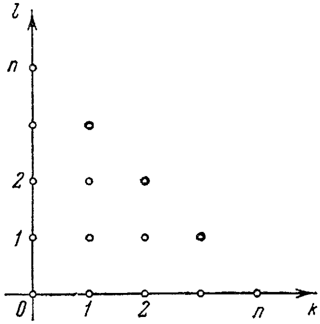


Fig. 79

For the system (3) to be determinate, it is necessary that the number of unknowns N be equal to the number of equations, whence, forming a "lattice of exponents" (Fig. 79), we obtain

$$N=(n+1)+n+\ldots+1=\frac{(n+1)(n+2)}{2}$$

A difficult and still open question is that of the most appropriate choice of lattice points for a given domain.

Another sufficiently general technique for computing a double integral may be indicated. Suppose the domain of integration is bounded by continuous single-valued curves

$$y=\varphi(x),\quad y=\psi(x)\quad (\varphi(x) \leqslant \psi(x))$$

and two vertical lines $x=a$, $x=b$ (Fig. 80).

Using familiar rules, set up the limits of integration in the double integral

$$I=\iint\limits_{(\sigma)} f(x,\ y)\,dx\,dy \tag{4}$$

to get

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \int\limits_{a}^{b} dx \int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy$$

Suppose

$$F(x) = \int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy \tag{5}$$

Then

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \int\limits_{a}^{b} F(x)\,dx \tag{6}$$

Applying one of the quadrature formulas to the single integral in the right-hand member of (6), we obtain

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{n} C_i F(x_i) \tag{7}$$

where $x_i \in [a, b]$ $(i = 1, 2, \ldots, n)$ and C_i are certain constant coefficients. In turn, the values

$$F(x_i) = \int\limits_{\varphi(x_i)}^{\psi(x_i)} f(x_i, y)\,dy$$

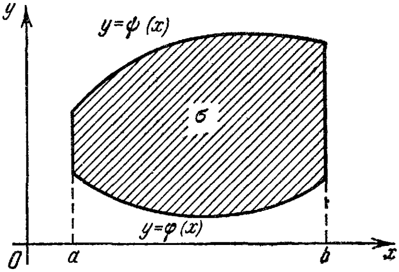


Fig. 80

may be found from certain quadrature formulas

$$F(x_i) = \sum_{j=1}^{m_i} B_{ij} f(x_i, y_j)$$

where B_{ij} are appropriate constants.

From formula (7) we derive

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{n} \sum_{j=1}^{m_i} C_i B_{ij} f(x_i, y_i) \tag{8}$$

where C_i and B_{ij} are known constants.

Geometrically, this method is equivalent to the computation of a volume I expressed by the integral (4) by means of cross-sections.

The general remarks made with respect to the computation of single integrals (see Sec. 16.10) remain valid, with appropriate modifications for cubature formulas of type (8).

*16.18 A CUBATURE FORMULA OF SIMPSON TYPE

To begin with, let the domain of integration be a rectangle:

$$R\{a \leqslant x \leqslant A, \quad b \leqslant y \leqslant B\}$$

(Fig. 81), whose sides are parallel to the coordinate axes. Halve each of the intervals $[a, A]$ and $[b, B]$ by the points

$$x_0 = a,\ x_1 = a + h,\ x_2 = a + 2h = A$$

and, respectively,

$$y_0 = b,\ y_1 = b + k,\ y_2 = b + 2k = B$$

where

$$h = \frac{A-a}{2}, \quad k = \frac{B-b}{2}$$

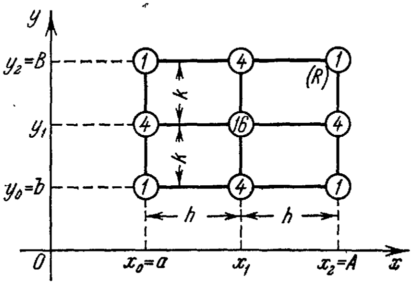


Fig. 81

In all, we thus obtain nine points (x_i, y_j) $(i, j = 0, 1, 2, \ldots, 9)$. We have

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \int\limits_a^A dx \int\limits_b^B f(x, y)\,dy \tag{1}$$

Now, computing the inner integral by Simpson's quadrature formula, we find

$$\iint f(x, y)\,dx\,dy = \int\limits_a^A dx \cdot \frac{k}{3}\,[f(x, y_0) + 4f(x, y_1) + f(x, y_2)] =$$

$$= \frac{k}{3}\left[\int\limits_a^A f(x, y_0)\,dx + 4\int\limits_a^A f(x, y_1)\,dx + \int\limits_a^A f(x, y_2)\,dx\right]$$

Again applying Simpson's formula to each integral, we get

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + 4f(x_1, y_0) + f(x_2, y_0)] +$$
$$+ 4\,[f(x_0, y_1) + 4f(x_1, y_1) + f(x_2, y_1)] +$$
$$+ [f(x_0, y_2) + 4f(x_1, y_2) + f(x_2, y_2)]\}$$

or

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + f(x_2, y_0) + f(x_0, y_2) +$$
$$+ f(x_2, y_2)] + 4\,[f(x_1, y_0) + f(x_0, y_1) + \tag{2}$$
$$+ f(x_2, y_1) + f(x_1, y_2)] + 16f(x_1, y_1)\}$$

We will call (2) *Simpson's cubature formula.* Hence,

$$\iint_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}(\sigma_0 + 4\sigma_1 + 16\sigma_2) \tag{2'}$$

where σ_0 is the sum of the values of the integrand $f(x, y)$ at the **vertices** of the rectangle R, σ_1 is the sum of values of $f(x, y)$ at the **midpoints of the sides** of the rectangle R, $\sigma_2 = f(x_1, y_1)$ is the value of the function $f(x, y)$ in the **centre** of R. The multiplicities of these values are given in Fig. 81.

Example 1. Applying the Simpson cubature formula, evaluate the double integral [7]

$$I = \int_4^{4.4}\int_2^{2.6} \frac{dx\,dy}{xy}$$

Solution. We take

$$h = \frac{4.4-4}{2} = 0.2 \text{ and } k = \frac{2.6-2}{2} = 0.3$$

The corresponding values of the integrand $z = \frac{1}{xy}$ are listed in Table 75.

TABLE 75

COMPUTING A DOUBLE INTEGRAL BY SIMPSON'S FORMULA

y_j \ x_i	4.0	4.2	4.4
2.0	0.125000	0.119048	0.113636
2.3	0.108696	0.103520	0.0988142
2.6	0.096154	0.0915751	0.0874126

Applying the cubature formula (2), we obtain

$$I = \frac{0.2\cdot 0.3}{9}[(0.125000 + 0.113636 + 0.096154 + 0.0874126) + 4(0.119048 + 0.108696 + 0.0988142 + 0.0915751) + 16\cdot 0.103520] = 0.0250070$$

The exact value of this double integral is

$$\int_4^{4.4}\int_2^{2.6} \frac{dx\,dy}{xy} = \ln 1.3 \cdot \ln 1.1 = 0.0953108 \cdot 0.262364 = 0.0250061$$

Hence, the residual error is

$$\Delta = |0.025006 - 0.0250070| = 0.0000009 \approx 10^{-6}$$

If the dimensions of the rectangle $R\ \{a \leqslant x \leqslant A,\ b \leqslant y \leqslant B\}$ are great, increased accuracy of the cubature formula (2) is obtained by partitioning the domain R into a system of rectangles and applying Simpson's cubature formula to each.

Suppose that the sides of the rectangle R are divided into n and m equal parts, respectively; the result will be a relatively coarse net nm of rectangles (in Fig. 82, the vertices of these rectangles are indicated by the large circles). We then subdivide each of these rectangles into four equal parts. The vertices of this fine net of rectangles will represent the lattice points M_{ij} of the cubature formula.

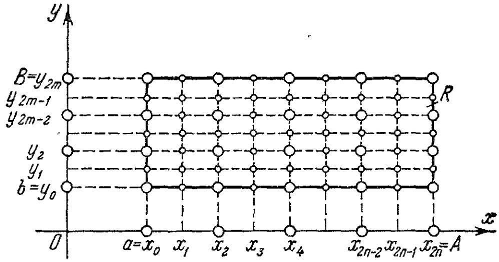


Fig. 82

Let

$$h = \frac{A-a}{2n}$$

and

$$k = \frac{B-b}{2m}$$

Then the lattice of points will have the following coordinates

$$x_i = x_0 + ih \qquad (x_0 = a; \quad i = 0,\ 1,\ 2,\ \ldots,\ 2n)$$

and

$$y_j = y_0 + jk \qquad (y_0 = b, \quad j = 0,\ 1,\ 2,\ \ldots,\ 2m)$$

For the sake of brevity, we introduce the notation

$$f(x_i,\ y_i) = f_{ij}$$

Applying formula (2) to each of the rectangles of the coarse net, we have (Fig. 82)

$$\iint\limits_{(R)} f(x,\ y)\,dx\,dy = \frac{hk}{9} \sum_{i=0}^{n} \sum_{j=0}^{m} [(f_{2i,\ 2j} + f_{2i+2,\ 2j} + f_{2i+2,\ 2j+2} + f_{2i,\ 2j+2}) + 4\,(f_{2i+1,\ 2j} + f_{2i+2,\ 2j+1} + f_{2i+1,\ 2j+2} + f_{2i,\ 2j+1}) + 16 f_{2i+1,\ 2j+1}]$$

Whence, collecting like terms, we finally get

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\sum_{i=0}^{2n}\sum_{j=0}^{2m}\lambda_{ij}f_{ij} \tag{3}$$

where the coefficients λ_{ij} are the corresponding elements of the matrix

$$\Lambda = \begin{bmatrix} 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ . & . & . & . & . & . & . & . & . & . \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \end{bmatrix}$$

If the domain of integration σ is curvilinear, then we construct a rectangle $R \supset \sigma$ whose sides are parallel to the coordinate axes

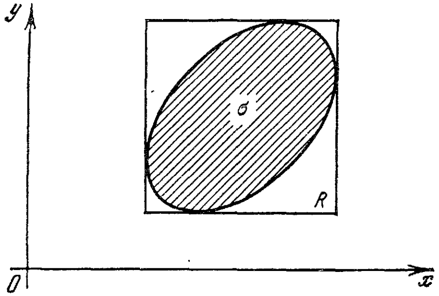


Fig. 83

(Fig. 83). We consider the auxiliary function

$$f^*(x, y) = \begin{cases} f(x, y) & \text{if } (x, y) \in \sigma, \\ 0 & \text{if } (x, y) \in R - \sigma \end{cases}$$

We then clearly have

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \iint\limits_{(R)} f^*(x, y)\,dx\,dy$$

This integral can be approximately evaluated by the general cubature formula (3).

REFERENCES FOR CHAPTER 16

[1] *Sh. E. Mikeladze, Numerical Methods of Mathematical Analysis*, 1953, Chapters XIII, XVIII (in Russian).
[2] *E. Milne, Numerical Calculus*, 1949, Chapter IV.
[3] *S. M. Nikolsky, Quadrature Formulas*, 1958 (in Russian).
[4] *A. Markov, Calculus of Finite Differences*, 1911, Chapter V (in Russian).
[5] *J. F. Steffensen, Interpolation*, 1927.
[6] *I. S. Berezin and N. P. Zhidkov, Computational Methods*, 1959, Volume 1, Chapter III (in Russian).
[7] *J. B. Scarborough, Numerical Mathematical Analysis*, 1955, Chapter VII.
[8] *A. N. Krylov, Lectures on Approximate Computations*, 1933, Chapter III (in Russian).
[9] *V. I. Krylov, Approximate Calculation of Integrals*, 1959 (in Russian).
[10] *M. G. Salvadori and M. L. Baron, Numerical Methods in Engineering*, 1952.
[11] *G. M. Fikhtengolts, Course of Differential and Integral Calculus*, 1948, Volume 2, Chapters IX, XIII (in Russian).

Chapter 17

THE MONTE CARLO METHOD

17.1 THE IDEA OF THE MONTE CARLO METHOD

The usual way of solving a problem is to indicate an *algorithm* (a sequence of operations) which yields the desired quantity f either exactly or to a specified accuracy. Namely, if we denote by $f_1, f_2, \ldots, f_n, \ldots$ the corresponding results of sequentially accumulating operations, then

$$f = \lim_{n \to \infty} f_n \tag{1}$$

and in the case of a finite number of operations, the process is terminated at some stage. Here the computation procedure is **strictly deterministic**: in the absence of mistakes, two different computors obtain the same result.

There are problems, however, for which it is practically impossible to construct an algorithm or the algorithm is prohibitively complicated. In such cases, one often resorts to modelling the mathematical or physical essence of the problem and to using the law of large numbers from probability theory. The estimates $f_1, f_2, \ldots, f_n, \ldots$ of the desired quantity f are obtained via a statistical treatment of material involving the results of certain *random trials* repeated many times. It is then required that the random quantity f_n converge in probability to the required quantity f as $n \longrightarrow \infty$ [1], [2]; that is for any $\varepsilon > 0$ we must have the limiting relation

$$\lim_{n \to \infty} \boldsymbol{P}(|f - f_n| < \varepsilon) = 1 \tag{2}$$

where $\boldsymbol{P}$ denotes the appropriate probability.

The choice of the quantity f_n is determined by the specific peculiarities of the problem. For example, one often interprets the sought-for quantity f as the probability of a random event (or, more generally, as the mathematical expectation of some random quantity). Then the frequency f_n of occurrence of the event for n random trials (or, respectively, the empirical mean of the values of a random quantity) may, under broad assumptions, be regarded as a probability estimate of the sought-for quantity. Other versions are also possible. It will be noted that in these cases the computational procedure is **nondeterministic** since it is determined by the results of random trials.

Methods of solving problems that employ random quantities are classed under the general heading of the *Monte Carlo method.* To be more precise, the *Monte Carlo method* [3], [4], [5], [6] embraces a collection of techniques which permit obtaining the solutions of mathematical or physical problems by means of repeated random trials. The estimates of a sought-for quantity are derived statistically and are of a probabilistic nature. In actual practice, random trials are replaced by the results of certain computations performed on *random numbers* (see Sec. 17.2).

The Monte Carlo approach came into effective use with the advent of high-speed electronic computers, since to obtain a sufficiently exact estimate of the desired quantity requires performing computations for a very large number of special cases with subsequent statistical treatment of a stupendous amount of numerical data. It must be pointed out that when using the Monte Carlo method there is no need to know the exact relationships between the given and the sought-for quantities of the problem; it suffices only to elucidate the set of conditions under which an appropriate event occurs. This circumstance makes it possible to use the Monte Carlo method in solving logical problems.

Of the mathematical problems to which the Monte Carlo method has been applied, we mention the following: solving systems of linear equations, matrix inversion, finding the eigenvalues and eigenvectors of a matrix, evaluating multiple integrals, solving the Dirichlet problem, solving functional equations of a variety of types, and so on. The Monte Carlo method is also successfully employed in the solution of problems of nuclear physics. It is well to bear in mind that even in the solution of one and the same problem the scheme for applying the method may be essentially different.

In this chapter we consider the computation of multiple integrals and the solution of systems of linear equations by the Monte Carlo method. Information concerning other types of problems may be found in the special literature (see, for example, [3], bibliography, and also [6]).

17.2 RANDOM NUMBERS

In actual applications of the Monte Carlo method, random trials are ordinarily replaced by sampling *random numbers.*

Definition 1. A *random* quantity is one whose values depend on the outcome of a random event.

A random quantity X is given by the distribution law

$$P(X < x) = \Phi(x)$$

where x is any real number and $\Phi(x)$ is a known function (the *distribution function*). The values of a random quantity are called *random numbers.*

Definition 2. If a random quantity has a given distribution law [1], [2] (uniform, normal, and so forth), then we will say that the appropriate random numbers *are distributed by this law.*

Suppose the numbers $x_1, x_2, \ldots, x_n, \ldots$ are the values of one and the same random quantity X under independent trials with recurrent conditions. Then the *sequence of random numbers*

$$\{x_n\} \tag{1}$$

is called a *random* sequence with the appropriate distribution law. In what follows we will, as a rule, consider random sequences (1) *uniformly distributed* on the unit interval $0 \leqslant x \leqslant 1$. If (a, b) is any interval[1] extracted from $[0, 1]$ and $\nu_n = \nu_n(a, b)$ is the number of elements in a finite subsequence $x_1, x_2, \ldots, x_n$ belonging to (a, b), then for the uniformly distributed sequence (1) we have the limiting relation

$$\lim_{n\to\infty} \frac{\nu_n(a, b)}{n} = b - a \tag{2}$$

which means that the *limiting relative frequency of the sequence* $\{x_n\}$ *uniformly distributed on* $[0, 1]$ *is, for each subinterval* (a, b), *equal to the length of the subinterval with probability* 1.

If the random sequence $\{x_n\}$ is uniformly distributed on the interval $[0, 1]$, then the linear transformation

$$y_n = A + (B - A)x_n \qquad (n = 1, 2, \ldots) \tag{3}$$

where A and B are given numbers, reduces to the random sequence $\{y_n\}$ uniformly distributed on the interval $[A, B]$. Generally, having a random sequence $\{x_n\}$ uniformly distributed on the interval $[0, 1]$, we can construct a random sequence $\{y_n\}$ with a specified distribution law $\Phi(y)$. Namely, suppose

$$\Phi(y) = \int_{-\infty}^{y} \varphi(t)\, dt$$

---

[1] By agreement, the endpoints a and b may or may not be included in the interval (a, b).

is the appropriate distribution function,[1] where $\varphi(t)$ is the *probability density.*

For the sake of simplicity, we assume that the function

$$x = \Phi(y)$$

is continuous and strictly monotonic (Fig. 84). Then, determining y_n from the equation

$$x_n = \Phi(y_n) \qquad (n = 1, 2, \ldots),$$

we obtain for each x_n the random sequence $\{y_n\}$ having the specified distribution law $\Phi(y)$. Namely, by the mode of construction

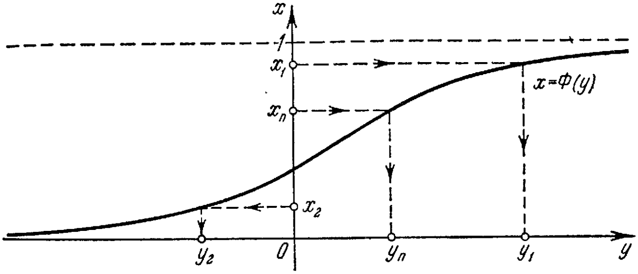


Fig. 84

the following limiting relation will hold, with probability 1, for the sequence $\{y_n\}$:

$$\lim_{n \to \infty} \frac{\tilde{\nu}_n(a, b)}{n} \int_a^b \varphi(y)\, dy \tag{4}$$

where $\tilde{\nu}_n(a, b)$ is the number of elements of the finite sequence $y_1, \ldots, y_n$ belonging to the arbitrary interval (a, b).

In particular, putting

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

we obtain by this method a canonical normally distributed (Gaussian) random sequence $\{y_n\}$ which corresponds to a random quantity Y with mathematical expectation $\boldsymbol{M}Y = 0$ and variance $\boldsymbol{D}Y = 1$. The linear transformation

$$z_n = \sigma y_n + c \qquad (n = 1, 2, \ldots)$$

---

[1] If y_n $(n = 1, 2, \ldots)$ lie in the finite interval $A \leqslant y \leqslant B$, then we as usual assume $\varphi(y) = 0$ for $y \notin [A, B]$.

yields a normally distributed random sequence $\{z_n\}$, which corresponds to the random quantity Z for which the expectation $MZ = c$ and the variance $DZ = \sigma^2$.

17.3 WAYS OF GENERATING RANDOM NUMBERS

Random numbers may be generated by using the results of random physical processes (say, dice, roullete wheels, scintillations in a Geiger-Müller counter, noise generated in electrical transmission systems, etc.). There are also available tables of random numbers (see, for example, [7], [8]).

Strictly speaking, when we use mechanical devices to obtain random numbers one is not completely sure that we are dealing with random events with a specified distribution of probabilities. Such material is therefore tested statistically for randomness. In this sense, tables of random numbers are a more reliable source, since they have already been tested for randomness. However, the use of tables of random numbers for solving problems on digital computers often involves great inconvenience [9].

Monte Carlo solutions ordinarily require a very large number of random numbers. For practical purposes, these numbers are most conveniently obtained by means of *special randomizing devices* which are hooked up to the computer. The operation of these devices is regulated by random physical processes (such as radioactive decay, noise in electronic tubes, etc.) [9].

Since the generation of random numbers that meet a given theoretical model is an extremely delicate and involved process, one often confines himself to obtaining so-called *pseudorandom numbers*, which largely resemble the corresponding random numbers. Rather involved mathematical algorithms are required for the production of pseudorandom numbers. In the sequel, the term "random number" will include both random and pseudorandom numbers if the distinction between them is not essential.

We now give some simple techniques for obtaining random numbers (in the broad meaning) uniformly distributed on the interval [0, 1]. For the sake of simplicity, we assume that these numbers are pure decimal fractions with a fixed number of decimals, say s (s-*digit* decimal fractions); that is, such that can be written as

$$x = \frac{\alpha_1}{10} + \frac{\alpha_2}{10^2} + \dots + \frac{\alpha_s}{10^s} \tag{1}$$

where α_i $(i = 1, 2, \dots, s)$ are the digits of this number which take the values 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

In order to compile a table of random numbers of type (1) uniformly distributed on [0, 1], it is sufficient to indicate the methods of obtaining the digits α_i and to meet the following conditions:

(a) α_i is a *random sampling* taken from the set of numbers from 0 to 9, all indicated values being equally probable and mutually independent;

(b) the choice of a digit α_i does not in the least affect the choice of the next digit α_{i+1}.

A sampling of this type is performed s times in order to obtain an s-digit random number.

The mode of choice which meets conditions (a) and (b) can actually be realized in many ways. We consider several.

1. Place in an urn ten identical balls numbered from 0 to 9. The balls are drawn from the urn in succession and the numbers α are recorded. After each extraction, the ball is returned to the urn and all the balls in the urn are mixed before each new extraction.

2. Two dice are thrown at one time. If n_1 and n_2 are the sums obtained (n_1, $n_2 = 1, 2, 3, 4, 5, 6$) of the first and second die, respectively (the dice must be distinguishable), then the digit α of the random number is taken equal to the remainder left after dividing the sum $6(n_1-1)+n_2$ by 10, where $n_1 < 6$, that is, α is a nonnegative integer less than 10 that satisfies the congruence[1]

$$6(n_1-1)+n_2 \equiv \alpha \pmod{10} \tag{2}$$

If $n_1 = 6$, the dice are thrown again. From formula (2) it follows that the digit α can with equal probability assume any value from 0 to 9 (see [7]).

3. An s-digit integer is squared and the middle s digits are extracted; then the process is repeated. If s is sufficiently great, say $s \geqslant 10$, then the extracted digits may be taken as sets of elements of s-digit pseudorandom numbers [3].

To obtain a sequence of pseudorandom numbers, one can also multiply a multidigit number by a fixed multiplier and extract middle digits or square a multidigit number and reduce modulo some sufficiently large prime.

4. A pseudorandom sequence $\{x_n\}$ is generated by means of the process [10]

$$x_n = 2^{-42}u_n$$

where

$$u_0 = 1, \quad u_{n+1} \equiv 5^{17}u_n \pmod{2^{42}}$$

---

[1] The notation $a \equiv b \pmod{k}$ (a, b, k integers) states that the difference $a-b$ is exactly divisible by k.

5. Use is made of the decimal expansion of a positive irrational number

$$\omega=\beta_0,\ \beta_1\beta_2\ \ldots\ \beta_s\ \ldots=\beta_0+(\omega)$$

where β_0 is the integral part of the number ω and (ω) is its fractional part.

To obtain a random sequence $\{x_n\}$ we assume

$$x_n=(n\omega)\qquad (n=1,\ 2,\ \ldots)$$

If a random sequence consisting of s-digit numbers is required, then one confines himself to the appropriate digits in the numbers $(n\omega)$.

To solve certain problems, one needs several random sequences:

$$\{x_n^{(1)}\},\ \{x_n^{(2)}\},\ \ldots,\ \{x_n^{(m)}\}$$

In this case, one chooses m linearly independent positive irrationalities $\omega_1,\ \omega_2,\ \ldots,\ \omega_m$ and puts

$$x_n^{(k)}=(n\omega_k)\qquad (k=1,\ 2,\ \ldots,\ m;\ n=1,\ 2,\ \ldots)$$

It is also possible to take one uniformly distributed random sequence $\{x_n\}$ and perform m samplings from it:

$$\begin{gathered}\{x_1,\ x_{m+1},\ x_{2m+1},\ \ldots\},\\ \{x_2,\ x_{m+2},\ \ldots\ x_{2m+2},\ \ldots\},\\ \cdots\cdots\cdots\cdots\cdots\\ \{x_m,\ x_{2m},\ x_{3m},\ \ldots\}\end{gathered}$$

taking numbers after every m elements instead of in succession. We will then obviously have m uniformly distributed subsequences.

TABLE 76

RANDOM NUMBERS UNIFORMLY DISTRIBUTED ON [0, 1]

0.57705	0.35483	0.11578	0.65339	0.66674
0.71618	0.09393	0.93045	0.93382	0.99279
0.73710	0.30304	0.93011	0.05758	0.24202
0.70131	0.55186	0.42844	0.00336	0.94010
0.16961	0.64003	0.52906	0.88222	0.60981
0.53324	0.20514	0.09461	0.98585	0.13094
0.43166	0.00188	0.99602	0.52103	0.35193
0.26275	0.55709	0.69962	0.91827	0.64560
0.05926	0.86977	0.31311	0.07069	0.64559
0.66289	0.31303	0.27004	0.13928	0.68008

Tables of random numbers are compiled by these and other methods. These tables ordinarily give random decimal digits, from which it is easy to construct random numbers of a specific size. By way of an example, we give a portion of a table (see [7]) of five-digit random numbers (Table 76).

17.4 MONTE CARLO EVALUATION OF MULTIPLE INTEGRALS

Let the function

$$y=f(x_1, x_2, \ldots, x_m)$$

be continuous in a closed bounded domain S and let it be required to evaluate the m-fold integral

$$I=\iint\limits_{(S)}\ldots\int f(x_1, x_2, \ldots, x_m)\,dx_1dx_2 \ldots dx_m \tag{1}$$

Geometrically, the number I is an $(m+1)$-dimensional volume of a right cylindroid[1] in the space $Ox_1x_2 \ldots x_my$ constructed on the base S and bounded above by the given surface $y=f(\boldsymbol{x})$, where $\boldsymbol{x}=(x_1, x_2, \ldots, x_m)$ (Fig. 85).

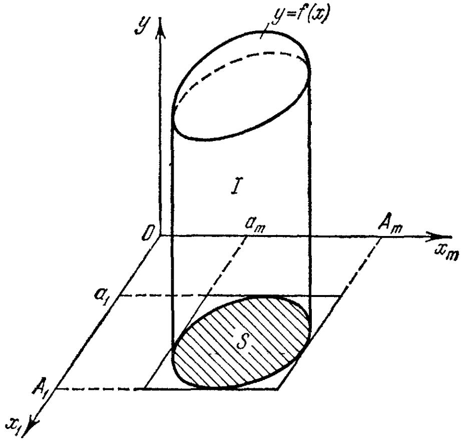


Fig. 85

We transform the integral (1) so that the new region of integration lies entirely within the unit m-dimensional hypercube. Let the domain S be located in the m-dimensional parallelepiped

$$a_i \leqslant x_i \leqslant A_i \tag{2}$$
$$(i=1, 2, \ldots, m)$$

---

[1] More precisely, an algebraic volume in which it is assumed that the parts of the cylindroid located above the hyperplane $Ox_1x_2 \ldots x_m$ are of positive measure, those below that hyperplane, of negative measure.

We make the change of variables

$$x_i = a_i + (A_i - a_i)\xi_i \qquad (3)$$
$$(i = 1,\ 2,\ \ldots,\ m)$$

It is then clear that the m-dimensional parallelepiped (2) will be transformed into the unit m-dimensional hypercube

$$0 \leqslant \xi_i \leqslant 1 \qquad (i = 1,\ 2,\ \ldots,\ m) \qquad (4)$$

and, consequently, the new region of integration σ which is found by the ordinary rules will lie entirely inside this hypercube (Fig. 86).

Computing the Jacobian of the transformation, we have

$$\frac{D(x_1,\ x_2,\ \ldots,\ x_m)}{D(\xi_1,\ \xi_2,\ \ldots,\ \xi_m)} = \begin{vmatrix} A_1 - a_1 & 0 & \ldots & 0 \\ 0 & A_2 - a_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & A_m - a_m \end{vmatrix} =$$
$$= (A_1 - a_1)(A_2 - a_2)\ldots(A_m - a_m)$$

Thus

$$I = \int\int\underset{(\sigma)}{\ldots}\int F(\xi_1,\ \xi_2,\ \ldots,\ \xi_m)\, d\xi_1\, d\xi_2 \ldots d\xi_m \qquad (5)$$

where

$$F(\xi_1,\ \xi_2,\ \ldots,\ \xi_m) = (A_1 - a_1)(A_2 - a_2) \ldots (A_m - a_m) \times$$
$$\times f(a_1 + (A - a_1)\xi_2,\ a_2 + (A_1 - a_2)\xi_2,\ \ldots,\ a_m + (A_m - a_m)\xi_m)$$

Introducing the notation

$$\xi = (\xi_1,\ \xi_2,\ \ldots,\ \xi_m)$$

and

$$d\sigma = d\xi_1\, d\xi_2 \ \ldots\ d\xi_m$$

we can write (5) more compactly as

$$I = \int\int\underset{(\sigma)}{\ldots}\int F(\xi)\, d\sigma \qquad (5')$$

We will give two ways for evaluating integral (5′) by the method of random trials.

First method: Choose on the interval [0, 1] m uniformly distributed independent sequences of random numbers:

$$\begin{array}{l} \xi_1^{(1)},\ \xi_2^{(1)},\ \ldots,\ \xi_n^{(1)},\ \ldots, \\ \xi_1^{(2)},\ \xi_2^{(2)},\ \ldots,\ \xi_n^{(2)},\ \ldots, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \xi_1^{(m)},\ \xi_2^{(m)},\ \ldots,\ \xi_n^{(m)},\ \ldots \end{array}$$

The points $M_i(\xi_i^{(1)},\ \xi_i^{(2)},\ \ldots,\ \xi_i^{(m)})$ $(i = 1,\ 2,\ \ldots)$ may be regarded as random. Choosing a sufficiently large number N of points M_1, M_2, ..., M_N, we check to see which belong to the domain σ *(first category)* and which do not *(second category)*. Let (Fig. 87)

(1) $$M_i \in \sigma \quad \text{for} \quad i = 1, 2, \ldots, n \tag{6}$$

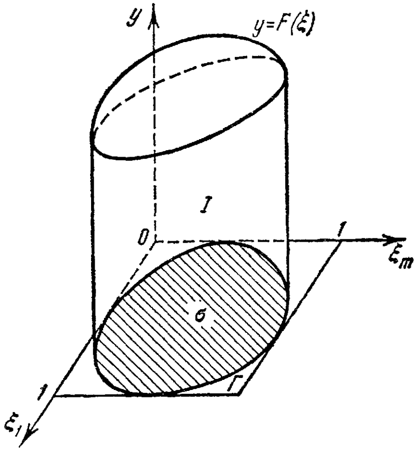


Fig. 86

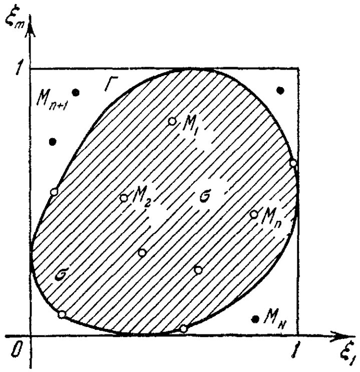


Fig. 87

and

(2) $$M_i \notin \sigma \quad \text{for} \quad i = n+1, n+2, \ldots, N \tag{6'}$$

(for the sake of convenience we change the numbering of the points here!). With regard to the boundary Γ of the domain σ, it is necessary beforehand to specify which boundary points, or what part of them, are to be regarded as belonging to σ and which as not belonging to this domain. This is of no particular significance in the general case for a smooth boundary Γ; but in certain cases it must be settled with regard for the concrete situation.

Taking a sufficiently large number n of points $M_i \in \sigma$, we put, approximately,

$$y_{av} = \frac{1}{n} \sum_{i=1}^{n} F(M_i)$$

whence the desired integral is given by the formula

$$I = y_{av}\sigma = \frac{\sigma}{n} \sum_{i=1}^{n} F(M_i) \tag{7}$$

where σ is understood to mean an m-dimensional volume in the domain of integration σ. If it is difficult to compute the volume of σ, then we can take

$$\sigma \approx \frac{n}{N}$$

whence

$$I \approx \frac{1}{N} \sum_{i=1}^{n} F(M_i)$$

In the particular case where σ is the unit cube ($\sigma = 1$), verification is unnecessary; thus $n = N$ and we simply have

$$I = \frac{1}{N} \sum_{i=1}^{N} F(M_i)$$

In verifying the conditions (6) and (6′), one ordinarily proceeds from the analytic representation of the boundary Γ of the domain σ. In the simplest case, if the surface Γ is given by the equation

$$\varphi(\xi) = 0 \tag{8}$$

where for $\varphi(\xi) < 0$ the point $\xi \in \sigma$ and for $\varphi(\xi) > 0$ the point $\xi \notin \sigma$, we have: (1) if $\varphi(M_i) < 0$, then the point M_i is of the first category and (2) if $\varphi(M_i) > 0$, then the point M_i is of the second category. The points M_i for which $\varphi(M_i) = 0$ are classed in the first or second category by convention. Note that equation (8) may be replaced by any equivalent equation. This sometimes greatly facilitates computation. Thus, for instance, the inequality for a circle

$$x^2 + y^2 - x - y + \frac{1}{4} \leqslant 0$$

is more conveniently replaced by the equivalent inequality

$$\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 \leqslant \frac{1}{4}$$

since the latter one is more easily verified.

If the domain σ is standard and given by the inequalities

$$\left.\begin{aligned} &\underline{\xi}_1 \leqslant \xi_1 \leqslant \bar{\xi}_1, \\ &\underline{\xi}_2(\xi_1) \leqslant \xi_2 < \bar{\xi}_2(\xi_1), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &\underline{\xi}_m(\xi_1, \ldots, \xi_{m-1}) \leqslant \xi_m \leqslant \bar{\xi}_m(\xi_1, \ldots, \xi_{m-1}) \end{aligned}\right\} \tag{9}$$

then to determine the membership of a random point $M(\xi_1, \xi_2, \ldots \xi_m)$ in the first or second category, one verifies the validity of these inequalities.

TABLE 77

SCHEME FOR DETERMINING MEMBERSHIP OF A RANDOM POINT $M(\xi_1, \ldots, \xi_m)$ IN THE STANDARD DOMAIN (9)

ξ_1	$\underline{\xi}_1$	$\overline{\xi}_1$	ε_1	ξ_2	$\underline{\xi}_2$	$\overline{\xi}_2$	ε_2

...	ξ_m	$\underline{\xi}_m$	$\overline{\xi}_m$	ε_m	ε	y

This is conveniently done by the scheme given in Table 77. Here

$$\varepsilon_i = \begin{cases} 1 & \text{if } \xi_i \in |\underline{\xi}_i, \overline{\xi}_i| \\ 0 & \text{if } \xi_i \notin |\underline{\xi}_i, \overline{\xi}_i| \end{cases}$$

$(i = 1, 2, \ldots, m)$ and $\varepsilon = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_m$. Clearly

$$\text{if } \varepsilon = 1, \text{ then } M \in \sigma,$$
$$\text{if } \varepsilon = 0, \text{ then } M \notin \sigma$$

Note that if $\varepsilon_j = 0\ (j < m)$, then the subsequent values $\varepsilon_{j+1}, \ldots, \varepsilon_m$ need not be computed since they do not affect the final result. The value of the function $y = F(M)$ is computed only for those points M for which $\varepsilon = 1$. Then formula (7) is used to evaluate the integral I.

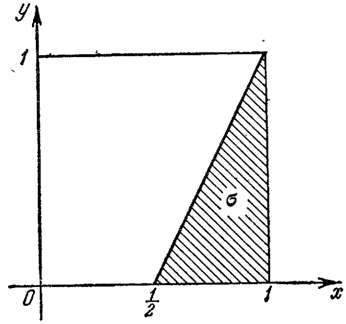


Fig. 88

Example. Give an approximate evaluation, by the Monte Carlo method, of the integral

$$I = \iint\limits_{(\sigma)} (x^2 + y^2)\, dx\, dy \tag{10}$$

where the domain σ of integration is defined by the inequalities

$$\frac{1}{2} \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 2x - 1 \tag{σ}$$

(Fig. 88).

Solution. The integral (10) is given in reduced form, that is the domain σ of integration is located in the unit square

$$0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1$$

To solve this problem, we take advantage of the table of random numbers (Table 76), regarding each succeeding pair of numbers in the table as the corresponding coordinates x and y of the random point $M(x, y)$. Since the computation is of an illustrative nature, we confine ourselves to $N=20$ random points; for the sake of simplicity, we round their coordinates to three decimal places. The results of the computations are listed in Table 78 where we put

$$\underline{x} = \frac{1}{2}, \qquad \overline{x} = 1,$$

$$\underline{y}(x) = 0, \qquad \overline{y}(x) = 2x - 1, \qquad z = x^2 + y^2$$

TABLE 78

EVALUATING THE DOUBLE INTEGRAL (10) BY THE MONTE CARLO METHOD

x	$\underline{x}$	$\overline{x}$	ε_1	y	$\underline{y}(x)$	$\overline{y}(x)$	ε_2	ε	z
0.577	0.500	1.000	1	0.716	0	0.154	0	0	
0.737	0.500	1.000	1	0.701	0	0.474	0	0	
0.170	0.500	1.000	0	0.533				0	
0.432	0.500	1.000	0	0.263				0	
0.059	0.500	1.000	0	0.663				0	
0.355	0.500	1.000	0	0.094				0	
0.303	0.500	1.000	0	0.552				0	
0.640	0.500	1.000	1	0.205	0	0.280	1	1	0.452
0.002	0.500	1.000	0	0.557				0	
0.870	0.500	1.000	1	0.323	0	0.740	1	1	0.855
0.116	0.500	1.000	0	0.930				0	
0.930	0.500	1.000	1	0.428	0	0.860	1	1	1.048
0.529	0.500	1.000	1	0.095	0	0.058	0	0	
0.996	0.500	1.000	1	0.700	0	0.992	1	1	1.482
0.313	0.500	1.000	0	0.270				0	
0.653	0.500	1.000	1	0.934	0	0.306	0	0	
0.058	0.500	1.000	0	0.003				0	
0.882	0.500	1.000	1	0.986	0	0.764	0	0	
0.521	0.500	1.000	1	0.918	0	0.042	0	0	
0.071	0.500	1.000	0	0.239				0	
$\sum$								4	3.837

Thus

$$z_{av}=\frac{1}{4}\cdot 3.837=0.96$$

and, hence, by formula (7), noting that $\sigma=\frac{1}{4}$, we have

$$I=z_{av}\cdot\sigma=0.96\cdot\frac{1}{4}=0.24 \tag{11}$$

If we take, approximately,

$$\sigma\approx\frac{n}{N}=\frac{4}{20}=\frac{1}{5}$$

then we get

$$I\approx 0.96\cdot\frac{1}{5}=0.19$$

Note that the true value of the integral is

$$I=\frac{7}{32}\approx 0.22$$

and so the relative error of the result (11) is

$$\delta=\frac{0.24-0.22}{0.22}\approx 9\%$$

The number of points here, $N=20$, is of course insufficient for the statistical regularities to exhibit themselves properly; nevertheless, a result satisfactory enough for a rough orientation has been obtained.

Second method. If the function $F(\xi)=F(\xi_1, \xi_2, \ldots, \xi_m)$ is nonnegative, then the integral (5) may be regarded as the volume of a solid V in the $(m+1)$-dimensional space $O\xi_1\xi_2\ldots\xi_m y$; that is,

$$I=\iint\limits_{(V)}\cdots\int d\xi_1 d\xi_2\ldots d\xi_m\,dy \tag{12}$$

where the domain of integration V is defined by the conditions

$$\xi=(\xi_1, \xi_2, \ldots, \xi_m)\in\sigma,\quad 0\leqslant y\leqslant F(\xi)$$

Let

$$0\leqslant F(\xi)\leqslant B \tag{13}$$

Introducing in (12) the new variable

$$\eta=\frac{1}{B}y \tag{14}$$

we get

$$I=B\iint\limits_{(v)}\cdots\int d\xi_1\,d\xi_2\ldots d\xi_m\,d\eta$$

where the new domain v is a cylindroid in the space $O\xi_1\xi_2 \ldots \xi_m\eta$ constructed on the region σ and bounded below by the hyperplane $\eta = 0$ and from above by the hyperplane

$$\eta = \frac{1}{B} F(\xi)$$

(Fig. 89). By virtue of inequality (13), the volume v lies entirely within the $(m+1)$-dimensional hypercube

$$0 \leqslant \xi_i \leqslant 1$$
$$(i = 1, 2, \ldots, m)$$
$$0 \leqslant \eta < 1$$

Now take, on $[0, 1]$, $m+1$ uniformly distributed independent random sequences

$$\{\xi_i^{(1)}\}, \{\xi_i^{(2)}\}, \ldots, \{\xi_i^{(m)}\}, \{\eta_i\}$$

whose corresponding elements will be regarded as the coordinates of the random points

$$M_i\{\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(m)}, \eta_i\}$$
$$(i = 1, 2, \ldots)$$

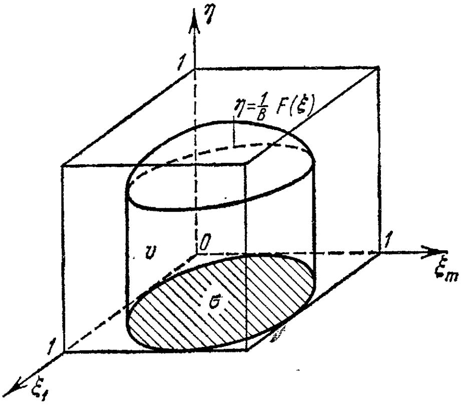


Fig. 89

of space $O\xi_1\xi_2 \ldots \xi_m\eta$. If n random points out of the total number of N belong to the volume v, and $N-n$ do not belong to this volume, then for a sufficiently large number N we can approximately put

$$I \approx B \cdot \frac{n}{N} \tag{15}$$

Thus,

$$I = B \cdot P(M \in v)$$

where the point M can occupy the positions $M_1, M_2, \ldots, M_N$ with the same probability. The validity of the relation

$$M \in v$$

is verified in the same way as in the first method. Note that if σ is a unit hypercube $0 \leqslant \xi_i \leqslant 1$ $(i = 1, 2, \ldots, m)$, then for the point $M_i(\xi_i^{(1)}, \ldots, \xi_i^{(m)}, \eta_i)$, all the coordinates of which are assumed to belong to the unit interval $[0, 1]$, it is sufficient to verify only the validity of the relation

$$\eta_i \leqslant \frac{1}{B} F(\xi_1^{(1)}, \xi_2^{(2)}, \ldots, \xi_m^{(m)})$$

Let us now consider the general case when

$$F(\xi)=F(\xi_1, \xi_2, \ldots, \xi_m)$$

is an alternating function. Suppose

$$-b \leqslant F(\xi) \leqslant B \tag{16}$$

where b and B are nonnegative numbers. Put

$$F(\xi)=-b+(B+b)\tilde{F}(\xi)$$

and we have

$$\iint\limits_{(\sigma)} \ldots \int F(\xi)\, d\sigma = -b\sigma+(B+b)\iint\limits_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma$$

where the function $\eta=\tilde{F}(\xi)$ satisfies, by inequality (16), the inequalities

$$0 \leqslant \tilde{F}(\xi) \leqslant 1$$

The integral

$$\iint\limits_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma = \iint\limits_{(\tilde{v})} \ldots \int d\sigma\, d\eta$$

may be evaluated by the method indicated above.

To estimate the accuracy of the approximate equation[1]

$$I_0=\iint\limits_{(v)} \ldots \int d\sigma\, d\eta = \boldsymbol{P}(M \in v) \approx \frac{n}{N} \tag{17}$$

first suppose that we have to do with ideal random uniformly distributed sequences of points M_i $(i=1, 2, \ldots)$, whose coordinates lie in the unit interval $[0, 1]$.

On the basis of the Bernoulli theorem and applying the Chebyshev inequality, we have

$$\boldsymbol{P}\left(\left|\frac{n}{N}-I_0\right|<\varepsilon\right) \geqslant 1-\frac{I_0(1-I_0)}{\varepsilon^2 N} \geqslant 1-\frac{1}{4\varepsilon^2 N} \tag{18}$$

Given, for a specified ε, the guarantee probability

$$P\left(\left|\frac{n}{N}-I_0\right|<\varepsilon\right) \geqslant 1-\delta \tag{19}$$

we get from inequality (18) that the condition (19) is definitely met if

$$\frac{1}{4\varepsilon^2 N}=\delta \tag{20}$$

From this we derive

$$\varepsilon=\frac{1}{2\sqrt{\delta N}} \tag{21}$$

---

[1] The factor B is inessential.

Thus, the accuracy of the estimate

$$I_0 \approx \frac{n}{N}$$

for a given maximum probability is inversely proportional to the square root of the number of trials, or $\varepsilon = O\left(\frac{1}{\sqrt{N}}\right)$. This circumstance causes the relatively slow convergence of the Monte Carlo method; for example, in order to reduce the error of the result 10-fold, the number of trials must be increased 100-fold! If the accuracy of the estimate ε and the guarantee probability $1-\delta$ are given, then from formula (20) we derive the necessary number of trials

$$N = \frac{1}{4\varepsilon^2\delta} \tag{22}$$

For example, for $\varepsilon = 0.001$ and $\delta = 0.01$ we have

$$N = 25{,}000{,}000$$

The estimate (22) is exaggerated and may be substantially improved!

We note one important circumstance: the number of trials N is independent of the dimensionality of the integral I_0 and therefore the Monte Carlo method is used to advantage in computing multiple integrals of high dimensionality, where the use of ordinary cubature formulas encounters appreciable difficulties. For example, in order to approximate in the ordinary way a 10-fold integral extended over a unit volume with spacing $h = 0.1$, one requires a sum containing roughly 10^{10} terms!

In practical Monte Carlo evaluation of multiple integrals, one ordinarily uses s-digit uniformly distributed random sequences. In this case, the faction $\frac{n}{N}$ will, if N is great, be close not to the true volume I_0 but to a certain fictitious volume I_0', which approximately represents the relative measure of the number of points M with coordinates of the form

$$\xi_i = \frac{k_i}{10^s}, \qquad \eta = \frac{k}{10^s} \tag{23}$$

$$(i = 1, 2, \ldots, m; \qquad k_i,\ k = 0, 1, 2, \ldots, 10^s)$$

that fall in volume v (cf. Sec. 17.3); strictly speaking, I_0' depends on whether the boundary points belong to the volume v or not. The total error of the result is estimated in the following manner (see [2]):

$$\left|\frac{n}{N} - I_0\right| \leqslant |I_1' - I_0| + \left|I_0' - \frac{n}{N}\right| \tag{24}$$

The first term $|I_0'-I_0|$ of the right member of (24) is the *usual computational error* obtained when replacing the integral I_0 by a sum corresponding to a partition of the volume v into elementary cubic cells whose vertices belong to the grid (23). The magnitude of this error may be evaluated by means of the inequality

$$|I_0'-I_0| \leqslant \overline{v}-\underline{v} \tag{25}$$

where $\overline{v}$ is the upper sum [in our case, for integral (17), it is simply the volume of the circumscribed step-like solid] and $\underline{v}$ is the lower sum (that is, the volume of the inscribed step-like solid). The magnitude of the error $|I_0'-I_0|$ depends essentially on the number of digits s in the random numbers, and if the boundary of the solid v is piecewise smooth, then this error can be made arbitrarily small for sufficiently large s. The inconvenience due to increasing the number of digits lies in the resulting increase in the amount of labour, since the computations then involve additional digits. The second term $\left|I_0'-\frac{n}{N}\right|$ in the right-hand member of inequality (24) is called the *sampling error* and may be estimated probabilistically by means of the Bernoulli theorem, as indicated above.

*17.5 SOLVING SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS BY THE MONTE CARLO METHOD

Let us consider the linear system

$$\sum_{j=1}^{n} a_{ij}x_j=b_i \qquad (i=1, \ldots, n) \tag{1}$$

In some way we reduce system (1) to the special form

$$x_i=\sum_{j=1}^{n} \alpha_{ij}x_j+\beta_i \qquad (i=1, \ldots, n) \tag{2}$$

Introducing the matrix $\alpha=[\alpha_{ij}]$ and the vectors

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \beta=\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}.$$

we can write system (2) in the matrix-vector form

$$\boldsymbol{x}=\alpha\boldsymbol{x}+\beta \tag{2'}$$

We will assume that all eigenvalues of the matrix α are less than unity in modulus. In particular, it suffices to consider that some canonical norm of the matrix α obeys the inequality

$$\|\alpha\| < 1 \tag{3}$$

Then system (2′) has a unique solution which can be found by the method of iteration (Sec. 8.10).

We select a set of multipliers v_{ij} so that the numbers p_{ij} defined by the equations

$$\alpha_{ij} = p_{ij} v_{ij} \quad (i, j = 1, \ldots, n) \tag{4}$$

satisfy the following conditions:

(1) $p_{ij} \geqslant 0$, and $p_{ij} > 0$ for $\alpha_{ij} \neq 0$,

(2) $\sum\limits_{j=1}^{n} p_{ij} < 1 \quad (i = 1, \ldots, n)$.

Let

$$p_{i,\, n+1} = 1 - \sum_{j=1}^{n} p_{ij} \qquad (i = 1, \ldots, n)$$

Besides, we agree to put

$$p_{n+1,\, j} = 0 \quad \text{for} \quad j < n+1$$

and

$$p_{n+1,\, n+1} = 1$$

Let us consider a particle performing a random walk and possessing a finite number of possible and incompatible states

$$S_1,\ S_2,\ \ldots,\ S_n,\ S_{n+1}$$

This particle is such that it passes from state S_i to state S_j with probability p_{ij} $(i, j = 1, \ldots, n+1)$, irrespective of previous states and with total indeterminacy relative to future states. The state $S_{n+1} = \Gamma$ (a "boundary" or "absorbing barrier") is **special** and corresponds to a complete stop of the particle, since, by virtue of the condition $p_{n+1,\, j} = 0$ $(j = 1, \ldots, n)$, transitions from the state S_{n+1} to the state S_j are impossible with probability 1 for $j < n+1$. Thus, the process of a random walk ceases as soon as a particle reaches the boundary Γ. The foregoing change of states is ordinarily called a *discrete Markov chain* (more precisely, *simple homogeneous*) with a finite number of states [2]. The numbers p_{ij} are called *transition probabilities*, and the matrix

$$\Pi = \begin{bmatrix} p_{11} & \cdots & p_{1n} & p_{1,\, n+1} \\ \cdot & \cdot & \cdot & \cdot \\ p_{n1} & \cdots & p_{nn} & p_{n,\, n+1} \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

is the transition probability matrix of the states $\{S_i\}$ *(chain law)*.

Let S_i be a fixed state different from the boundary state $(i < n+1)$. We consider the random walk of a particle which starts out from the given state $S_i = S_{i_0}$ and, after a series of intermediate states S_{i_1}, S_{i_2}, ..., S_{i_m}, terminates on the boundary $S_{i_{m+1}} = \mathbf{\Gamma}$. Thus, S_{i_m} $(m \geqslant 0)$ is the state of the particle immediately preceding its emergence at the boundary. The totality of states

$$T_i = \{S_{i_0}, S_{i_1}, \ldots, S_{i_m}, S_{i_{m+1}}\} \tag{5}$$

will for brevity be called a *trajectory (path)*. Let X_i be a random quantity dependent on the random trajectories T_i starting with the state S_i (the *functional of the trajectory* T_i) and assuming on the trajectory (5) the value

$$\xi(T_i) = \beta_{i_0} + v_{i_0 i_1}\beta_{i_1} + v_{i_0 i_1} v_{i_1 i_2}\beta_{i_2} + \ldots + v_{i_0 i_1} \ldots v_{i_{m-1} i_m}\beta_{i_m} \tag{6}$$

where β_j $(j = i_0, i_1, \ldots, i_m)$ are the corresponding constant terms of the reduced system (2).

In particular, if $v_{ij} = 1$, we simply have

$$\xi(T_i) = \beta_{i_0} + \beta_{i_1} + \ldots + \beta_{i_m} \tag{6'}$$

By the product rule for probabilities, the trajectory T_i and, hence, the value $\xi(T_i)$, occurs with probability

$$\mathbf{P}(T_i) = p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_m i_{m+1}} \tag{7}$$

where $i_0 = i$ and $i_{m+1} = n+1$.

Theorem. *The mathematical expectations*

$$MX_i = x_i \qquad (i = 1, 2, \ldots, n)$$

are the roots of system (2).

Proof. The trajectories T_i which start out from the state S_i may be partitioned into $n+1$ categories

$$\begin{aligned} T_{i_1} &= \{S_i, S_1, S_{i_2}, \ldots\}, \\ T_{i_2} &= \{S_i, S_2, S_{i_2}, \ldots\}, \\ &\ldots\ldots\ldots\ldots \\ T_{i_n} &= \{S_i, S_n, S_{i_2}, \ldots\}, \\ T_{i,\,n+1} &= \{S_i, S_{n+1}\} \end{aligned}$$

depending on the first step; that is, a particle which initiates a random walk from state S_i can, at the first step, either pass into state S_1 or into state S_2, and so on, and upon completion of a certain number of steps terminate the random walk at the boundary.

If a particle has the trajectory

$$T_{ij} = \{S_i, S_j, S_{i_2}, \ldots, S_{i_m}, S_{i_{m+1}} = \mathbf{\Gamma}\}$$

where $j \neq n+1$, then the random quantity X_i will, by (6), take on the value

$$\begin{aligned}\xi(T_{ij}) &= \beta_i + v_{ij}\beta_j + v_{ij}v_{ji_2}\beta_{i_2} + \ldots + \\ &\ldots + v_{ij}v_{ji_2} \ldots v_{i_{m-1}i_m}\beta_{i_m} = \beta_i + v_{ij}(\beta_j + v_{ji_2}\beta_{i_2} + \\ &+ \ldots + v_{ji_2} \ldots v_{i_{m-1}i_m}\beta_{i_m}) = \beta_i + v_{ij}\xi(T_j) \qquad (8)\end{aligned}$$

where T_j is some trajectory with initial state S_j.

When the particle reaches the boundary Γ immediately, that is, when the path is of the form $T_{i,\,n+1} = \{S_{i,\,n+1}\}$, then

$$\xi(T_{i,\,n+1}) = \beta_i. \qquad (8')$$

The probability that the trajectory T_i is a trajectory of type T_{ij} is obviously equal to p_{ij}.

By the definition of mathematical expectation, we have

$$MX_i = \sum_{T_i} \xi(T_i)\,P(T_i) = \sum_j \sum_{T_{ij}} \xi(T_{ij})\,P(T_{ij})$$

If $j < n+1$, then the trajectory T_{ij} consists of the interval (S_i, S_j) and some trajectory T_j. For this reason, $P(T_{ij}) = p_{ij}P(T_j)$. For $j = n+1$ we have

$$\xi(T_{i,\,n+1}) = \beta_i \quad \text{and} \quad P(T_{i,\,n+1}) = p_{i,\,n+1}$$

Moreover, since each trajectory T_{ij} is for $j < n+1$ associated in a one-to-one manner with the trajectory T_j (and vice versa) summation over the trajectories T_{ij} for $j = 1, 2, \ldots, n$ may be replaced by summation over the trajectories T_j.

Whence, having regard for formula (8), we get

$$MX_i = \sum_{j=1}^{n} \sum_{T_j} [\beta_i + v_{ij}\xi(T_j)] \cdot p_{ij}P(T_j) + \beta_i p_{i,\,n+1}$$

or

$$MX_i = \sum_{j=1}^{n} p_{ij}v_{ij} \sum_{T_j} \xi(T_j)P(T_j) + \beta_i \left[\sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\,n+1}\right]$$

But, clearly,

$$\sum_{T_j} \xi(T_j)\,P(T_j) = MX_j \qquad (j = 1, 2, \ldots, n)$$

Moreover, $\sum_{T_j} P(T_j) = 1$ and

$$\sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\,n+1} = \sum_{j=1}^{n+1} p_{ij} = 1$$

Hence

$$MX_i = \sum_{j=1}^{n} \alpha_{ij}MX_j + \beta_i \qquad (i = 1, \ldots, n)$$

where $\alpha_{ij} = p_{ij}v_{ij}$.

The proof of the theorem is complete.

Note. In proving this theorem we assumed that the mathematical expectations

$$x_i = MX_i \qquad (i = 1, \ldots, n)$$

exist. It can be proved that when condition (3) is met the random quantities X_i have finite expectations.

From the theorem just proved it follows that the roots of system (2) may be regarded as the mathematical expectations of the random quantities $X_1, \ldots, X_n$. For an experimental determination of the quantity $x_i = MX_i$ one organizes N random walks with random trajectories $T_i^{(k)}$ $(k = 1, \ldots, N)$ with initial state S_i and each time one records the value $\xi(T_i^{(k)})$ of the random quantity X_i. Suppose that the trials are independent and the quantity X_i has a bounded variance. Then, by virtue of Chebyshev's theorem [1], [2], for N sufficiently large, the following inequality will hold true with a probability arbitrarily close to unity:

$$\left| x_i - \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}) \right| < \varepsilon$$

where ε is the given limiting error. Thus, the roots of the system (2) can approximately be determined from the formulas

$$x_i \approx \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}) \tag{9}$$

In particular, this method can be used to invert a matrix of the form

$$A = E - \alpha \tag{10}$$

where $\|\alpha\| < 1$ and $E = [\delta_{ij}]$ is the unit matrix. To do this, note that the elements of the inverse matrix

$$A^{-1} = [x_{ij}]$$

are the roots of the linear system

$$\sum_{k=1}^{n} (\delta_{ik} - \alpha_{ik}) x_{kj} = \delta_{ij} \quad (i, j = 1, \ldots, n)$$

whence we find that the elements of each column

$$x_{1j}, \ldots, x_{nj} \qquad (j = 1, \ldots, n)$$

of the matrix A^{-1} are determined from the linear subsystem

$$x_{ij} = \sum_{k=1}^{n} \alpha_{ik} x_{kj} + \delta_{ij} \qquad (i = 1, \ldots, n) \tag{11}$$

On the basis of the foregoing, starting from the state $S_i = S_{i_0}$, for fixed j, we obtain the random quantity X_{ij} with values

$$\xi_j(T_i) = \delta_{i_0 j} + \delta_{i_1 j} v_{i_0 i_1} + \ldots + \delta_{i_m j} v_{i_0 i_1} \ldots v_{i_{m-1} i_m}$$

where $T_i = \{S_{i_0}, S_{i_1}, \ldots, S_{i_m} S_{i_{m+1}} = \Gamma\}$ and the numbers v_{ij} are such that p_{ij}, determined from the equations $\alpha_{ij} = p_{ij} v_{ij}$, are transition probabilities from the state S_i to the state S_j. The expectations $MX_{ij} = x_{ij}$ yield the desired elements of the matrix A^{-1}.

Let us now show, practically, how to organize a random walk of a particle with given transition probabilities p_{ij}. For the sake of simplicity, we assume that p_{ij} are decimal fractions with common denominator 10^s (s natural):

$$p_{i1} = \frac{t_{i1}}{10^s}, \quad p_{i2} = \frac{t_{i2}}{10^s}, \quad \ldots, \quad p_{i,\,n+1} = \frac{t_{i,\,n+1}}{10^s}$$

where $t_{i1}, t_{i2}, \ldots, t_{i,\,n+1}$ are nonnegative integers, and

$$t_{i1} + t_{i2} + \ldots + t_{i,\,n+1} = 10^s \qquad (i = 1, 2, \ldots, n)$$

We consider a particle with initial state S_i. Let $\{x\}$ be s-digit numbers less than unity uniformly distributed on the interval $[0, 1]$; for example, the elements of a table of random numbers. Let us generate the random number x. If it happens that the inequality

$$0 \leqslant x < \frac{t_{i1}}{10^s}$$

holds true, then we will take it that the particle moves from state S_i to state S_1. Furthermore, if

$$\frac{t_{i1}}{10^s} \leqslant x < \frac{t_{i1} + t_{i2}}{10^s}$$

then we assume that the particle moves from S_i to S_2. The other transitions are defined in a similar manner. In particular, a particle hits the boundary $S_{n+1} = \Gamma$ if the random number x is such that

$$\frac{t_{i1} + \ldots + t_{in}}{10^s} \leqslant x < \frac{t_{i1} + \ldots + t_{in} + t_{i,\,n+1}}{10^s} = 1$$

On the basis of the given convention it is clear that the number of favourable cases for the transitions $S_i \to S_j$ ($j = 1, 2, \ldots, n+1$) are proportional to the respective numbers

$$t_{i1}, t_{i2}, \ldots, t_{i,\,n+1}$$

and these cases are equally probable. Therefore the transition probabilities

$$P(S_i \to S_j) = \frac{t_{ij}}{10^s} = p_{ij} \qquad (i = 1, \ldots, n; \quad j = 1, \ldots, n+1)$$

Extracting a sequence of random numbers and taking the above rule for guidance, we get the random walk of a particle with fixed initial state and given transition probabilities. To obtain the required accuracy of the roots (in a probabilistic sense) one must consider a sufficiently large number of independent random walks.

Example. Solve the following system of equations by the Monte Carlo method:

$$\left.\begin{aligned} x_1 &= 0.1x_1 + 0.2x_2 + 0.7, \\ x_2 &= 0.2x_1 - 0.3x_2 + 1.1 \end{aligned}\right\} \tag{12}$$

Solution. We can put

$$\begin{aligned} v_{11} &= 1, & v_{12} &= 1, \\ v_{21} &= 1, & v_{22} &= -1 \end{aligned}$$

whence the transition probability matrix is

$$\Pi = \left[\begin{array}{cc:c} 0.1 & 0.2 & 0.7 \\ 0.2 & 0.3 & 0.5 \\ \hdashline 0 & 0 & 1 \end{array}\right]$$

where the elements of the first row are, respectively, the probabilities of transition from state S_1 to states S_1, S_2 and $S_3 = \Gamma$, and the elements of the second row are those from state S_2 to states S_1, S_2 and S_3, the "fringe" corresponding to the boundary Γ.

Since the elements of the matrix Π are multiples of 0.1, one can use single-digit random numbers whose digits are recruited from some randon sequence, say, are elements of the random numbers of Table 76 (Sec. 17.3).

The results obtained for 20 random walks with initial state S_1 are listed in Table 79. The random number x ensured the transitions of the states in accord with the following instructions:

I. For initial state S_1:

(1) if $0 \leqslant x \leqslant 0.1$, then $S_1 \to S_1$,
(2) if $0.1 \leqslant x < 0.3$, then $S_1 \to S_2$,
(3) if $0.3 \leqslant x < 1$, then $S_1 \to \Gamma$

II. For initial state S_2:

(1) if $0 \leqslant x < 0.2$, then $S_2 \to S_1$,
(2) if $0.2 \leqslant x < 0.5$, then $S_2 \to S_2$,
(3) if $0.5 \leqslant x < 1$, then $S_2 \to \Gamma$.

The values of the random quantity X_1 computed from formula (6) are listed in the last column of Table 79. From this

$$x_1 = MX_1 \approx \frac{1}{20}(20 \cdot 0.7 + 0.7 + 4 \cdot 1.1) = 0.96$$

TABLE 79
FINDING THE UNKNOWN x_1 OF SYSTEM (12)
BY THE MONTE CARLO METHOD

No.	Random number x.	Random-walk trajectory	Value of random quantity X_1
1	0.5	$S_1 \longrightarrow \Gamma$	0.7
2	0.7	$S_1 \longrightarrow \Gamma$	0.7
3	0.7	$S_1 \longrightarrow \Gamma$	0.7
4	0.0 0.5	$S_1 \longrightarrow S_1 \longrightarrow \Gamma$	$0.7+0.7$
5	0.7	$S_1 \longrightarrow \Gamma$	0.7
6	0.1 0.6	$S_1 \longrightarrow S_2 \longrightarrow \Gamma$	$0.7+1.1$
7	0.1 0.8	$S_1 \longrightarrow S_2 \longrightarrow \Gamma$	$0.7+1.1$
8	0.7	$S_1 \longrightarrow \Gamma$	0.7
9	0.3	$S_1 \longrightarrow \Gamma$	0.7
10	0.7	$S_1 \longrightarrow \Gamma$	0.7
11	0.1 0.0 0.7	$S_1 \longrightarrow S_2 \longrightarrow S_1 \longrightarrow \Gamma$	$0.7+1.1+0.7$
12	0.0 0.1 0.3 0.1 0.1 0.6	$S_1 \longrightarrow S_1 \longrightarrow S_2 \longrightarrow S_2 \longrightarrow$ $\longrightarrow S_1 \longrightarrow S_2 \longrightarrow \Gamma$	$0.7+0.7+1.1-$ $-1.1-0.7-1.1$
13	0.9	$S_1 \longrightarrow \Gamma$	0.7
14	0.6	$S_1 \longrightarrow \Gamma$	0.7
15	0.1 0.5	$S_1 \longrightarrow S_2 \longrightarrow \Gamma$	$0.7+1.1$
16	0.3	$S_1 \longrightarrow \Gamma$	0.7
17	0.3	$S_1 \longrightarrow \Gamma$	0.7
18	0.2 0.4 0.4 0.3 0.1 0.6	$S_1 \longrightarrow S_2 \longrightarrow S_2 \longrightarrow S_2 \longrightarrow$ $\longrightarrow S_2 \longrightarrow S_1 \longrightarrow \Gamma$	$0.7+1.1-1.1+$ $+1.1-1.1-0.7$
19	0.6	$S_1 \longrightarrow \Gamma$	0.7
20	0.2 0.6	$S_1 \longrightarrow S_2 \longrightarrow \Gamma$	$0.7+1.1$
		Σ	$21 \cdot 0.7+4 \cdot 1.1$

The unknown x_2 is computed in similar fashion.

Note that the exact roots of system (12) are $x_1 = 1$ and $x_2 = 1$.

Other methods are also employed for solving algebraic linear equations by the Monte Carlo method [11].

REFERENCES FOR CHAPTER 17

[1] *E. S. Ventsel, The Theory of Probability*, 1958, Chapters I-VI (in Russian).
[2] *B. V. Gnedenko, The Theory of Probability*, 1969, Chapters 1-6 (translated from the Russian).
[3] *A. S. Householder, Principles of Numerical Analysis*, 1953, Chapter 8.
[4] *W. E. Milne, Numerical Solution of Differential Equations*, 1953, Appendix C.
[5] *Yu. A. Shreider, A Method of Statistical Trials* (Monte Carlo), 1956 (in Russian).
[6] *Edwin F. Beckenbach (editor), Modern Mathematics for the Engineer*, First Series, 1956; Chapter 12, Monte Carlo Methods by George W. Brown.
[7] *P. M. Morse and G. E. Kimball, Methods of Operations Research*, 1951, Chapter 6, Sec. 4.
[8] *M. Kadyrov, Tables of Random Numbers*, 1936 (in Russian).
[9] *A. I. Kitov and N. A. Krinitsky, Electronic Digital Computers and Programming*, 1959, Chapter VIII (in Russian).
[10] *P. Davis, P. Rabinowitz, Some Monte Carlo Experiments in Computing Multiple Integrals*, 1956.
[11] *Yu. A. Shreider, Solution of Systems of Linear Algebraic Equations by the Monte Carlo Method*, 1958 (in Russian).

COMPLETE LIST OF REFERENCES

Beckenbach, Edwin F. (editor), *Modern Mathematics for the Engineer*, First Series, McGraw-Hill Book Company, New York, 1956 (for Chapters 8, 13, 17).

Berezin, I. S. and Zhidkov, N. P., *Computational Methods*, Fizmatgiz, 1959, in Russian (for Chapters 8, 12, 16).

Bezikovich, Ya. S., *Calculus of Finite Differences*, LGU, 1939, in Russian (for Chapter 6).

Bezikovich, Ya. S., *Approximate Computations*, Gostekhizdat, 1949, in Russian (for Chapters 1,4).

Bradis, V. M., *The Theory and Practice of Computations*, Moscow, Uchpedgiz, 1935, in Russian (for Chapter 14).

Bradis, V. M., Oral and Written Computing. Computational Aids, *Encyclopedia of Elementary Mathematics*, Book I, Gostekhizdat, 1951, in Russian (for Chapter 1).

Bulgakov, B. V., *Oscillations*, Gostekhizdat, 1954, in Russian (for Chapter 7).

But, E., *Numerical Methods*, Moscow, Fizmatgiz, 1959, in Russian (for Chapter 13).

Davis, P., Rabinowitz, P., Some Monte Carlo Experiments in Computing Multiple Integrals, *Math. Tables and Other Aids Comput.*, 1956, 10, No. 53, 1-8 (for Chapter 17).

Faddeyev, D. K. and Faddeyeva, V. N., *Computational Methods of Linear Algebra*, Fizmatgiz, 1960, in Russian (for Chapters 7, 8, 12).

Faddeyeva, V. N., *Computational Methods of Linear Algebra*, Gostekhizdat, 1950, in Russian (for Chapters 7, 8, 9, 10, 11, 12, 14).

Fikhtengolts, G. M., *Mathematics for Engineers*, GTTI, 1933, in Russian (for Chapter 1).

Fikhtengolts, G. M., *Principles of Mathematical Analysis*, Vol. I, Gostekhizdat, 1955, in Russian (for Chapter 2).

Fikhtengolts, G. M., *Course of Differential and Integral Calculus*, Nauka, 1969, in Russian (for Chapters 3, 16).

Fikhtengolts, G. M., *Course of Differential and Integral Calculus*, Gostekhizdat, 1957, in Russian (for Chapter 4).

Fikhtengolts, G. M., *Principles of Mathematical Analysis*, Vol. II, Gostekhizdat, 1956, in Russian (for Chapter 6).

Frazer, R. A., Duncan, W. J., Collar, A. R., *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, New York, The Macmillan Company, 1946 (for Chapter 7).

Fuks, B. A. and Shabat, B. V., *Functions of a Complex Variable*, Moscow-Leningrad, Gostekhizdat, 1949, in Russian (for Chapter 5).

Gantmacher, F. R., *The Theory of Matrices*, Moscow, Gostekhizdat, 1953, in Russian, or translated from the Russian, Chelsea New York, 1959 (for Chapter 10).

Gavurin, M. K., Application of Polynomials of Best Approximation to the Acceleration of Convergence of Iterative Processes, *Uspekhi Matem. Nauk*, 5:3 (**37**), 1950, in Russian (for Chapter 12).
Gelfand, I. M., *Lectures on Linear Algebra*, Moscow-Leningrad, Gostekhizdat, 1951, in Russian (for Chapters 10, 12).
Gelfond, A. O., *Calculus of Finite Differences*, Gostekhizdat, 1952, in Russian (for Chapters 4, 5, 6).
Gnedenko, B. V., *The Theory of Probability*, Moscow, MIR Publishers, 1969, translated from the Russian (for Chapter 17).
Goncharov, V. L., *The Theory of Interpolation and Approximation of Functions*, Moscow-Leningrad, GTTI, 1934, in Russian (for Chapter 14).
Grave, D., *Elements of Higher Algebra*, Kiev, 1914, in Russian (for Chapter 5).
Hildebrand, F. B., *Introduction to Numerical Analysis*, New York, McGraw-Hill Book Company, 1956 (for Chapter 5).
Householder, Alston S., *Principles of Numerical Analysis*, New York, McGraw-Hill Book Company, 1953 (for Chapters 10, 13, 17).
Kadyrov, M., *Tables of Random Numbers*, Tashkent, Publishing House of State University of Central Asia, 1936, in Russian (for Chapter 17).
Kagan, B. M. and Ter-Mikaelyan, T. M., *Solution of Engineering Problems on Automatic Digital Computers*, Moscow-Leningrad, Gosenergoizdat, 1958, in Russian (for Chapter 3).
Kantorovich, L. V., On Newton's Method, *Transactions of the Steklov Institute of Mathematics*, XXVIII, 1949, in Russian (for Chapters 4, 13).
Kantorovich, L. V., Krylov, V. I., *Approximate Methods of Higher Analysis*, Gostekhizdat, 3rd ed., 1949, in Russian (for Chapter 6).
Khinchin, A. Ya., *Continued Fractions*, Gostekhizdat, 1949, in Russian (for Chapter 2).
Khovansky, A. N., *Application of Continued Fractions and Their Generalizations to Problems of Approximate Analysis*, Gostekhizdat, 1956, in Russian (for Chapters 2, 3).
Kitov, A. I. and Krinitsky, N. A., *Electronic Digital Computers and Programming*, Moscow, Fizmatgiz, 1959, in Russian (for Chapter 17).
Krylov, A. N., *Lectures on Approximate Computations*, Leningrad, Academy of Sciences, USSR, 2nd ed., 1933, in Russian (for Chapters 1, 5, 16).
Krylov, A. N., *Lectures on Approximate Computations*, Gostekhizdat, 6th ed., 1954, in Russian (for Chapters 6, 15).
Krylov, V. I., *Approximate Calculation of Integrals*, Moscow, Fizmatgiz, 1959, in Russian (for Chapter 16).
Kurosh, A. G., *Course of Higher Algebra*, Moscow, MIR Publishers, 1972, translated from the Russian (for Chapters 5, 11, 12).
Lednev, N. A., (editor), Mathematical Practice Session Devoted to Computing Instruments and Machines, *Soviet Science*, 1959, Moscow, in Russian (for Chapter 14).
Lyapin, E. S., *Course of Higher Algebra*, Uchpedgiz, 1953, in Russian (for Chapter 7).
Lyusternik, L. A., *Transactions of the Steklov Institute of Mathematics*, **20**, 1947, page 49, in Russian (for Chapter 12).
Lyusternik, L. A., Abramov, A. A., Shestakov, V. I., Shura-Bura, M. R., *The Solution of Mathematical Problems on Automatic Digital Computers*, USSR Academy of Sciences Publishing House, 1952, in Russian (for Chapter 3).
Maltsev, A. I., *Principles of Linear Algebra*, Moscow-Leningrad, Gostekhizdat, 1948, in Russian (for Chapter 10).
Maltsev, A. I., *Principles of Linear Algebra*, Gostekhizdat, 2nd ed., 1956, in Russian (for Chapter 7).
Markov, A., *Calculus of Finite Differences*, Matezis, 2nd ed., 1911, in Russian (for Chapters 3, 6, 16).

Mikeladze, Sh. E., *Numerical Methods of Mathematical Analysis*, Moscow, Gostekhizdat, 1953, in Russian (for Chapters 15, 16).

Milne, W. E., *Numerical Calculus*, Princeton, New Jersey, Princeton University Press, 1949 (for Chapters 12, 14, 15, 16).

Milne, W. E., *Numerical Solution of Differential Equations*, New York, London, 1953 (for Chapters 13, 17).

Mlodzeyevsky, B. K., *Solution of Numerical Equations*, Moscow, GIZ, 1924, in Russian (for Chapter 5).

Morse, P. M. and Kimball, G. E., *Methods of Operations Research*, New York, 1951, Chapter 6, Sec. 4. (for Chapter 17).

Nikolsky, S. M., *Quadrature Formulas*, Moscow, Fizmatgiz, 1958, in Russian (for Chapter 16).

Ostrowski, Alexander M., *Sur la convergence et l'estimation des erreurs dans quelques procédés de résolution des équations numériques (Collection of Papers in Memory of D. A. Grave*, Moscow GITTL, 1940, p. 213) [Sbornik posvyashchonny pamyati D. A. Grave] (for Chapter 13).

Perron, O., *Die Lehre von den Kettenbrüchen*, Leipzig, Teubner, 1929 (for Chapter 2).

Remez, E. Ya., *General Computational Methods of Chebyshev Approximation*, Ukrainian Academy of Sciences Publishing House, 1957, in Russian (for Chapter 14).

Runge, Carl, *Graphische Methoden*, Teubner, Leipzig-Berlin, Zweite Auflage, 1919 (for Chapter 15).

Salekhov, G., *Calculation of Series*, Gostekhizdat, 1955, in Russian (for Chapter 6).

Salvadori, Mario G., and Baron, M. L., *Numerical Methods in Engineering*, New York, Prentice-Hall, Inc., 1952 (for Chapters 8, 16).

Scarborough, James B., *Numerical Mathematical Analysis*, Baltimore, Johns Hopkins Press, 3rd ed., 1955, (for Chapters 1, 4, 5, 8, 13, 14, 15, 16).

Schreier, O. und Sperner, E., Vorlesungen über Matrizen, Leipzig, Teubner, 1932 (for Chapter 7).

Shapiro, G. M., *Higher Algebra*, Moscow, GUPI, 4th ed., 1938, in Russian (for Chapter 5).

Shilov, G. E., *Introduction to the Theory of Linear Spaces*, Moscow-Leningrad, Gostekhizdat, 1952, in Russian (for Chapter 10).

Shreider, Yu. A., The Solution of Systems of Linear Algebraic Equations, *Doklady Akademii Nauk SSSR*, 5, 1951, in Russian (for Chapter 10).

Shreider, Yu. A., Solution of Systems of Linear Algebraic Equations by the Monte Carlo Method, Sbornik I: "*Problems in the Theory of Mathematical Machines*", Moscow, Fizmatgiz, 1958, in Russian (for Chapter 17).

Shreider, Yu. A., A Method of Statistical Trials (Monte Carlo), *Instrument Design Journal*, No. 7, 1956, in Russian (for Chapter 17).

Smirnov, V. I., *Course of Higher Mathematics*, Vol. 3, Moscow-Leningrad, GTTI, 1933, in Russian (for Chapter 11).

Smirnov, V. I., *Course of Higher Mathematics*, Vol. 1, Moscow, Gostekhizdat, 17th ed., 1957, in Russian (for Chapter 3).

Smolitsky, Kh. L., *Computational Mathematics (Lecture Notes)*, Leningrad, Mozhaisky LKVVIA, 1960, in Russian (for Chapter 8).

Steffensen, J. F., *Interpolation*, Baltimore, Williams and Wilkins, 1927 (for Chapter 16).

Tolstov, G. P., *Fourier Series*, Gostekhizdat, 1951, in Russian (for Chapter 6).

Tolstov, G. P., *Course of Mathematical Analysis*, Vol. I, Gostekhizdat, 1954, in Russian (for Chapter 4).

Tolstov, G. P., *Course of Mathematical Analysis*, Vol. II, Moscow, Gostekhizdat, 1957, in Russian (for Chapter 3).

Vallée Poussin, C. J., de la, *Cours d'Analyse Infinitesimale*, I (4th edition), 1921 (for Chapter 6).

Ventsel, E. S., *The Theory of Probability*, Moscow, Fizmatgiz, 1958, in Russian (for Chapter 17).

Ventsel, D. A., Ventsel, E. S. *Elements of the Theory of Approximate Computations*, Moscow, Publishing House of VVIA named after N. E. Zhukovsky, 1949, in Russian (for Chapters 1, 4, 13).

Wayland, Harold, Expansion of Determinantal Equations into Polynomial Form, *Quarterly of Applied Mathematics*, Vol. II, No. 4, Jan. 1945, pp. 277-306 (for Chapter 12).

Whittaker, E. T., and Robinson, G., *The Calculus of Observations*, London and Glasgow, 4th edition, 1944 (for Chapters 4, 7, 14).

Zaguskin, V. L., *Handbook on Numerical Methods of Solving Algebraic and Transcendental Equations*, Fizmatgiz, 1960, in Russian (for Chapter 5).

INDEX

Abel, N. H.
  Euler-Abel method 209
  Euler-Abel transformation 210
Abramov, A. A. 114, 676
absolute error 19
absolute value of a matrix 242
absorbing barrier 667
acceleration of convergence
  of Fourier trigonometric series by Krylov's method 217
  of numerical series 203
  of power series by Euler-Abel method 209
  of series 89
accumulation, method of 306, 640
accuracy
  in determination of arguments from a tabulated function 48
  estimation of 17
  of quadrature formulas 618
additive inverse of a matrix 232
adjoint of a matrix 236
algebra
  fundamental theorem of 162
  matrix 229
algebraic equations
  approximate solution of (special techniques for) 162ff
  bounds of real roots of 167
  general properties of 162ff
algebraic volume 656
algorithm
  Euclid's 165
  Hero's 107
alternating sums, method of 169
analysis, classical 5
analytic functions, computing values of 89
angle between two vectors 345
approximate differentiation 574ff
  formulas of, based on Newton's first interpolation formula 575
  formulas of, based on Stirling's formula 580
approximate integration 590ff
approximate numbers 19ff
approximate solution of algebraic equations, special techniques for 162ff
approximation
  of improper integrals 633
  major 19
  minor 19
  trigonometric 225
argument, principle of 166

back substitution (*see* direct procedure) 280
backward extrapolation 529
backward interpolation 529
Barlow's tables 113
Baron, M. L. 321, 648, 677
barrier, absorbing 667
bases (*see* basis)
basis (bases)
  biorthogonal 391
  normalized orthogonal 346
  orthogonal 346
  orthonormal 346
  of a space 340
    initial 341
Beckenbach, E. F. 321, 506, 674, 675
Berezin, I. S. 321, 458, 645, 675
Bernoulli method 198ff
Bernoulli numbers 99, 208, 625ff
  generating function of 627
Bernstein, S. N. 608
Bessel's formula for parabolic interpolation 535
Bessel's function of order zero 582
Bessel's inequality 216
Bessel's interpolation formula 534, 535
Bezikovich, Ya. S. 54, 161, 228, 675
bilinear expansion of a matrix 392
bilinear form of a matrix 384
biorthogonal bases 391
biorthogonality, conditions of 390
biorthogonality relations 390, 393
blocks of a matrix 256
bordered matrices 257
bordering, method of (matrices) 262
bounds
  method of 50
  of real roots of algebraic equations 167
Bradis, V. M. 54, 573, 675
Brown, G. W. 674
Budan
  theorem of Budan-Fourier 175, 176
Bulgakov, B. V. 272, 675
But, E. 506, 675

canonical convergents 59
canonical norm of a matrix 243
Cardan's formula 186
Cauchy inequality 245
Cauchy test 83
Cayley-Hamilton theorem 397
central derivatives, formulas for 580
central differences 530
central formulas 580
central-difference formulas 531
chain, discrete Markov 667
chain law 668
change-of-basis matrix 348
characteristic determinant 376, 565
characteristic equation 376
characteristic matrix 376
characteristic number 375, 377
characteristic polynomial 376, 382
characteristic root 375
Chebyshev, P. L. 554
Chebyshev polynomial 554
Chebyshev's quadrature formula 607ff
check
  final 17
  intermediate 17
check sums 281
chords, method of 122
class, periodicity 214
classical analysis 5
coefficients
  Cotes 594, 600
  Fourier 213
  Lagrangian 543
Collar, A. R. 272, 675
column vector 229
combination method 136
commutative matrices 234

complex roots
  one pair of 190
  two pairs of 194, 197
components of continued fractions, terms of 55
computation(s) (*see also* computing)
  of determinants 269
  double, method of 50
  in which errors are not taken into exact account 42
computation sheets 16
computational scheme of Danilevsky's method 416
computation work, rules of 15ff
computing (*see* also computation)
  analytic functions 89
  cube roots 112
  exponential functions 91
  forms 16
  functions 77ff
  hyperbolic functions 101
  logarithmic functions 95
  polynomials 77
  rational fractions 82
  reciprocals 104
  reciprocals of square roots 111
  square roots 107
  trigonometric functions 98
conditions
  of biorthogonality 390
  Hurwitz 406
  Sylvester 389
conformal partitioned matrices 257
continued fraction(s) 55
  components of 55
  conversion of to a simple fraction 56
  expanding functions into 72ff
  infinite 55
  n-component 55
  nonterminating 66
    convergent 66
    divergent 67
simple 56
standard 56
theory of 55ff
contraction mapping 487ff
convergence
  accelerating (by Lyusternik method) 453, 458
  of Fourier trigonometric series (acceleration of by Krylov method) 217
  of iteration processes
    first sufficient condition of 491
    second sufficient condition of 493
  of iteration processes for systems of linear equations 322ff, 394ff
    necessary and sufficient conditions for 398
    sufficient conditions for 322
  of matrix power series 394
  methods for effectively checking the conditions of 405
  of the Newton process 465, 469
    rapidity of 474
    stability of 478
  of numerical series, accelerating 203
  of power series, acceleration of by Euler-Abel method 209
  of series, acceleration of 89
  of Seidel process
    first sufficient condition for 327
    necessary and sufficient conditions for 400
    for a normal system 403
    second sufficient condition for 330
    third sufficient condition for 333
convergence theorem 214
convergent integral 633, 635
convergents 58ff
    canonical 59
    law of formation of 59
coordinates of a vector 336
    in a basis 341
correct digit 25
Cotes
  Newton-Cotes quadrature formulas 593
  Newton-Cotes formulas 599
Cotes coefficients 594, 600
Cramer's formulas 276
Cramer's rule 273
cubature, mechanical 590
cubature formulas 590, 641ff
  of Simpson type 644
  Simpson's 644
cube roots, computation of 112

Danilevsky, A. M.
  method of 412
    computation of eigenvector's by 420
  computational scheme of 416
  exceptional cases in 418
Davis, P. 674, 675
degenerate linear transformation 374
Del 497
delta
  Kronecker 230
  the operator Δ 508
derivatives
  central 580
  partial 588
Descartes' rule of signs 178
descent, steepest (*see* method of steepest descent), 496, 499
determinant(s) 230
  characteristic 376, 565
  computation of 269
  Gaussian method in computing 288
  secular 376, 565
    expansion of 410ff, 429
  Vandermonde 592, 614
diagonal difference table 511
diagonal matrix 229, 265
difference(s)
  central 530
  divided 554ff
  double (of higher order) 570
  error of 34
  finite 507ff
  lambda (λ-difference) 440
difference equation 198
difference table 510
  diagonal 511
  horizontal 511
differentiation
  approximate 574ff
  graphical 586
  numerical 583
differentiation operator 629
digit
  correct 25
  significant 24
dimensionality of a subspace 342
direct procedure (*see* forward substitution) 280
discrete Markov chain 667
distribution function 651
distributivity 344

divergent integral 633, 635
divided differences 554ff
  table of 556
double computation, method of 50, 622
double differences of higher order 570
double-entry table 569
Duncan, W. J. 272, 675

eigenvalue(s) 375, 377, 382, 445
  extremal property of 387
  finding (of a matrix) 410ff
  finding the first 436
  finding the numerically largest 430
  finding the second 439
eigenvector(s) 375, 382, 445ff
  computation of by Danilevsky's method 420
  computation of by Krylov's method 424
  finding (*see* eigenvalue) 410ff, 430, 439
element(s)
  of a matrix 229
  principal 287
Emde 524, 568
empirical formula 524
entries (of a matrix) 229
equal effects, principle of 45
equation(s)
  algebraic 162
  characteristic 376
  difference 198
  equivalent 119
  graphical solution of 119
  linear 273
  nonlinear (*see* systems of nonlinear equations) 459
  root of 115ff
  secular 376
  solution of 199
equivalent equations 119
equivalent matrices 268
error(s) 19
  absolute 19
  computations in which errors are not taken into exact account 42
  of a difference 35
  epsilon (ε-error) (propagation law of) 515
  general formula for 42
  initial 23
  limiting absolute 20
  limiting relative 21
  of method 22, 621
  of operation 23, 84
  of the problem 22
  of a product 37
  of a quotient 40
  relative (*see* relative error) 19, 21
  residual 23
  rounding 23
  sampling 666
  sources of 22ff
  of a sum 33
  theory of 42
error estimate, probability 52
escalator method 320
Euclid's algorithm 165
Euler, L. 58
Euler-Abel method 209
Euler-Abel transformation 210
Euler-Maclaurin formula 628, 630
Euler-Maclaurin summation formula 630
even-digit rule 26
exact methods 273
exact number 19
exhaustion, method of 443, 444
expansion
  bilinear (of a matrix) 392
  of secular determinants 410ff
    comparison of different methods of 429
  Stirling's 207
exponential functions 50, 91
  computing values of 91
extrapolation 519
  backward 529
  forward 529
  Richardson 622ff
extremal property of eigenvalues 387

factorial(s) (*see* generalized power) 517
  inverse 205
Faddeyev, D. K. 272, 321, 458, 657
Faddeyeva, V. N. 272, 321, 335, 393, 409, 458, 573, 675
Fikhtengolts, G. M. 54, 76, 114, 161, 228, 648, 675
final check 17
finite differences 507ff
finite-difference operator 628
form(s)
  bilinear (of a matrix) 384
  computing 16
  Frobenius standard 412
  quadratic (*see* quadratic form) 311
formula(s)
  of approximate differentiation
    based on Newton's first interpolation formula 575
    based on Stirling's formula 580
  Bessel's interpolation 534, 535
  Bessel's, for parabolic interpolation 535
  Cardan's 186
  central 580
  for central derivatives 580
  central interpolation 552
  central-difference 531
  Chebyshev's quadrature 607ff
  Cramer's 276
  cubature 590, 641ff
    of Simpson type 644
    Simpson's 644
  empirical 524
  Euler-Maclaurin 628, 630
  Euler-Maclaurin summation 630
  Gauss' quadrature 611, 614
  Gaussian interpolation 531, 532, 533
  general trapezoidal 601
  for interpolating to halves 536
  interpolation (*see* interpolation formulas)
  Lagrange's interpolation 539, 541
  Lambert's 100
  Markov's 566
  Newton-Cotes (of higher orders) 599
  Newton-Cotes quadrature 593
  Newton-Leibniz 590
  Newton's first interpolation 519, 522
  Newton's second interpolation formula 526, 527
  Newton's quadrature 599
  quadrature 590
    accuracy of 618

- of closed type 591
- Newton's 599
- of open type 591
- Simpson's (and its remainder term) 596
- Simpson's general 603
- Stirling's interpolation 533
- trapezoidal (and its remainder term) 595

Forsythe, G. E. 321
forward extrapolation 529
forward interpolation 529
forward substitution (*see* direct procedure) 280
Fourier
  theorem of Budan-Fourier 175, 176
Fourier coefficients 213
  estimates of 213
Fourier trigonometric series 213
fraction(s)
  continued (*see* continued fractions) 55
  rational (*see* rational fractions) 82
Frazer, R. A. 272, 675
Frobenius matrix 412, 415, 418
Frobenius standard form 412
Fuks, B. A. 202, 675
function(s)
  analytic 89
  Bessel's (of order zero) 582
  computing values of 77ff
  distribution 651
  exponential 50, 91
  hyperbolic 101
  interpolating 518
  interpolation of 507ff, 519
  iteration for approximating the values of 103ff
  jump 221
  logarithmic 95
  signum 86
  trigonometric 49, 98
  zero of 115
fundamental system of solutions 365
fundamental theorem of algebra 162

Gantmacher, F. R. 393, 675
Gauss' first interpolation formula 532
Gauss' quadrature formula 611, 614
Gauss' second interpolation formula 532, 533
Gaussian interpolation formulas 531, 532, 533
Gaussian method 277ff
  inversion of matrices by 290
  use of in computing determinants 288
Gaussian random sequence 652
Gavurin, M. K. 458, 676
  method of 458
Gelfand, I. M. 393, 458, 676
Gelfond, A. O. 161, 202, 228, 676
general formula for errors 42
general trapezoidal formula 601
generalized power 517
generate (*see* space, generated by vectors) 342
Gnedenko, B. V. 674, 676
Goncharov, V. L. 573, 676
gradient (of a function) 497
gradient method 496
Graeffe 182
  method of Lobachevsky-Graeffe 179, 182
graphical differentiation 586
graphical integration 639
graphical solution of equations 119
Grave, D. 202, 676

halving method 121
Hermitian symmetry 343
Hero's algorithm 107
Hero's process 107
Hildebrand, F. B. 201, 202, 676
horizontal difference table 511
Horner's scheme 77, 78
  generalized 80-82
Householder, A. S. 393, 506, 674, 676
Hua's theorem 179
Hurwitz conditions 405
Hurwitz theorem 406
hyperbolic cosine 101
hyperbolic functions 101
hyperbolic sine 101
hyperbolic tangent 102
hypercube, unit *m*-dimensional 656

identical transformation 375
improper integral(s) 633
  aproximation of 633
improving roots 284
inequality
  Bessel 216
  Cauchy 245
infinite continued fraction 55
initial basis of a space 341
initial error 23
integral(s)
  convergent 633, 635
  divergent 633, 635
  improper 633
  multiple 656
  probability 561
  proper 633
intergation
  approximate 590ff
  graphical 639
intermediate check 17
interpolating function 518
interpolating to halves, formula for 536
  interpolation 519
  backward 529
  forward 529
  of functions 507, 519
  of functions of two variables 567
  inverse (*see* inverse interpolation)
  method of 411
  in the narrow sense 519
  parabolic 522
  problem of (statement of) 518
interpolation formula(s)
  Bessel's 534, 535
    remainder term of 552
  central 552
  with constant interval (general) 536ff
  Gauss' first 532
  Gauss' second 533
  Gaussian 531
  Lagrange's 539, 541
    error estimate of 547
  Newton's 519
    error's estimates of 550
    for a function of two variables 571
    for unequally spaced values of the argument 556, 558
  Newton's first 519, 522

  Newton's second 526, 527
  Stirling's 533
    remainder term of 552
interpolation method, for expanding a secular determinant 565
interpolation points 518
  best choice of 553
interval (*see* spacing) 507
  variable 555
invariance of a linear subspace 379
inverse, additive (of a matrix) 232
inverse factorials 205
inverse interpolation
  for case of equally spaced points 559
  for case of unequally spaced points 562
  finding roots of an equation by 564
inverse matrix 236
  correcting elements of an approximate 316ff
  properties of 239
inverse problem of theory of errors 44
  second 48
inverse transformation 373
inversion of matrices 236
  by Gaussian method 290
  solution of systems of linear equations by 273
  isolation
  of roots 115
  of singularities by Kantorovich method 635
iteration (*see* linear system)
  for approximating the values of a function 103ff
  method of 138ff, 300, 302, 484ff
  process of, convergence of (*see* convergence of process of iteration) 491, 493
  of a vector 431
iteration processes
  convergence of (for systems of linear equations) 322ff, 394ff
  estimate of the error of approximations in 324
iterative methods (*see* methods, iterative, and methods of iteration)

Jacobi matrix 465, 470
Jacobian 156, 657
Jahnke 524, 568
Jump function 221

Kadyrov, M. 674
Kagan, B. M. 114, 676
Kantorovich, L. V. 161, 228, 465, 481, 506, 676 method of (for isolating singularities) 635
Kantorovich theorem 465
Khaletsky, scheme of 295, 297
Khinchin, A. Ya. 76, 676
Khovansky, A. N. 76, 114, 676
Kimball, G. E. 674, 677
Kitov, A. I. 674, 676
Krinitsky, N. A. 674, 676
Kronecker delta 230
Krylov, A. N. 54, 189, 202, 217, 228, 589, 648, 676
  method of 217, 225, 411, 421ff
    computation of eigenvectors by 424
Krylov, V. I. 225, 648, 676
Kummer transformation 203, 204
Kurosh, A. G. 202, 409, 458, 676

Lagrange, method of 169
Lagrange's interpolation formula 539, 541
Lagrange's theorem 168
Lagrangian coefficients 543
  computing 543
  computational scheme for 546, 547
lambda-difference 440
Lambert's formula 100
latent root 375
law
    chain 668
  propagation (of ε-error) 515
Lednev, N. A. 573, 676
Legendre polynomials 611
  properties of 611
Leibniz
  Newton-Leibniz formula 590
length of a vector 345
Leverrier, method of 411
limiting absolute error 20ff
limiting relative error 21
  tables for determining 30
Lin's method 201
linear dependence of vectors 337
linear equations, solving systems of 273ff
linear subspace 341
  invariance of 379
linear system
  normal 312
  reducing (to a form convenient for iteration) 307
linear transformations
  degenerate 374
  operations with 371
  singular 374
  of variables 367
linear vector spaces 336ff
  theory of 336ff
linear-transformation operator 369
linearly dependent vectors 337
linearly independent vectors 337
lines (of a matrix) 229
Lobachevsky, N. I. 182
Lobachevsky-Graeffe method 179, 182
  for case of complex roots 187, 192
  for case of real and distinct roots 184
logarithmic function 95
logarithms 49
loss of accuracy in subtraction 35
Lyapin, E. S. 272, 676
Lyusternik, L. A. 114, 453, 458, 676
  method of 453ff

Maclaurin
  Euler-Maclaurin formula 628, 630
Maclaurin's series 89
major approximation 19
Maltsev, A. I. 272, 393, 676
mapping, contraction 487ff
Markov, A. 114, 228, 648, 676
Markov chain, discrete 667
Markov's formula 566
matrices (*see* matrix)
matrix (matrices)
  absolute value of 242
  adjoint of 236

- bilinear expansion of 392
- bilinear form of 384
- bordered 257
- change-of-basis 348
- characteristic 376
- commutative 234
- conformal partitioned 257
- diagonal 229, 265
- difference of 231
- eigenvalues of 375
- eigenvectors of 375
- elementary transformations of 268
- equality of 230
- equivalent 268
- Frobenius 412, 415, 418
- inverse 236, 239
- inversion of (*see* inversion of matrices) 273
- Jacobi 465, 470
- limit of 249
- minor of 248
- modulus of 242
- multiplication of 232
  - by a scalar 231
- nonsingular 236
- norm of 242, 243
- nullity of 248
- operations involving 230
- orthogonal 350
  - method of 363
  - properties of 350
- orthogonalization of 351
- partitioned (*see* partitioned matrices)
- positive definite symmetric real 388
- powers of 240
- product of 231
- of quadratic form 311
- quasidiagonal 256
- rank of 248
- rational functions of 241
- real 389
- rectangular 229
- series of (*see* matrix series) 251
- similar 339, 380
- singular 236
- square 229
- sum of 231
- symmetric 235, 384
- trace of 377
- transformation 368
- transition probability 667
- transpose of 234
- triangular 265
- unit 230
- zero 230

matrix algebra 229

matrix inversion 236
- by partitioning 260
- using the coefficients of the characteristic polynomial of a matrix for 450

matrix power series, convergence of 394

matrix series, absolutely convergent 252

mechanical cubature 590

mechanical quadrature 590

mesh points 518

method(s)
- of A. A. Abramov 458
- of accumulation 306, 640
- of alternating sums 169
- Bernoulli's 198ff
- of bordering (matrices) 262
- of bounds 50
- of chords 122
- combination 136
- of *A. M. Danilevsky 412*
  - computation of eigenvectors by 420
  - computational scheme of 416
  - exceptional cases in 418
- *of double computation 50, 622*
- for effectively checking the conditions of convergence 405
- escalator 320
- *Euler-Abel 209*
- exact (in solving systems of linear equations) 273
- of exhaustion 443, 444
- Gaussian (*see* Gaussian methods) 277ff
- of M. K. Gavurin 458
- gradient 496
- halving 121
- interpolation 411
- of iteration 138ff, 300, 302, 484ff
- iterative (in solving systems of linear equations) 273
- of L. V. Kantorovich (for isolating singularities) 635
- of A. N. Krylov 217, 225, 411, 421ff
  - computation of eigenvectors by 424
- of Lagrange 169
- of Leverrier 411, 426
- Lin's 201
- of *Lobachevsky-Graeffe 179, 182*
- of L. A. Lyusternik 453ff
- modified Newton 135
- Monte Carlo 649ff
- Newton *127, 171, 459f*
  - modified 481ff
- nomographic 121
- of orthogonal matrices 363
- *orthogonalization (see orthogonalization methods)*
- of N. V. Paluver 201
- of power series 504
- *of principal elements 287*
- of proportional parts 122
- of Purcell 320
- of relaxation 313ff
- of Richardson 320
- of scalar products (for finding first eigenvalue of a real matrix) 436
- Seidel 309ff
- square-root 293
- of steepest descent 496, 499
  - for a system of linear equations 501
- Sturm 173
- of successive approximations 138
- of tangents 127
- of undetermined coefficients 411, 428ff

Mikeladze, Sh. E. 589, 648, 676

Milne, W. E. 458, 506, 573, 589, 648, 674, 677

Minor (of a matrix) 248

Minor approximation 19

Mlodzeyevsky, B. K. 202, 677

modified Newton method 135

modulus 242
- of a matrix 242
- Young's 44

Monte Carlo evaluation of multiple integrals 656

Monte Carlo method 649ff
- problems attacked by 650
- solving systems of linear algebraic equations by 666

Morrey, C. B. Jr. 506

Morse, P. M. 674, 677

multiple integrals, Monte Carlo evalu-

ation of 656
multiplicity of a root 162

nabla 497
negative
  of a matrix 232
  of a vector 336
negative definite quadratic form 312
Newton-Cotes formulas of higher orders 599
Newton-Cotes quadrature formulas 593
Newton-Leibniz formula 590
Newton's first interpolation formula 526, 522
Newton's interpolation polynomial 521
Newton's method 127, 171, 459ff
  for complex roots 157
  modified 135, 481ff
  for a system of two equations 156
Newton's process, convergence of 465, 469
  rapidity of 474
  stability of 478
Newton's quadrature formula 599
Newton's second interpolation formula 526, 527
Newton's theorem 171
Nikolsky, S. M. 592, 648, 677
nomographic methods 121
nonlinear equations (*see* systems of nonlinear equations) 459
nonsingular matrix 236
nonsingular transformation 374
nonterminating continued fractions 66
  convergent 67
  divergent 67
norm of a matrix 242, 243
  canonical 243
  k-norm, l-norm, m-norm 243
normal linear system 312
normal system 311
normalized orthogonal basis 346
notation
  powers-of-ten 18, 24
  scientific 18, 24
nullity of a matrix 248
number(s)
  approximate 19ff
  Bernoulli 99, 208, 625ff
  characteristic 375, 377
  exact 19
  pseudorandom 653
  random (*see* random numbers) 650
numerical differentiation, formulas for equally spaced points 583
numerical series 83
  approximation of sums of 83
  convergent 83

operator
  differentiation 629
  finite-difference 628
  linear-transformation 369
orthogonal basis 346
orthogonal matrices 350
  method of 363
  properties of 350
orthogonal systems of vectors 345
orthogonal vectors 345
orthogonality 345
orthogonalization
  of columns 358
  of matrices 351
  of rows 361
orthogonalization methods, application of to solution of systems of linear equations 358
orthonormal basis 346
Ostrowski, A. 465, 506, 677
Ostrowski's theorem 159, 465

Paluver, N. V. 201
  method of 201
parabolic interpolation 522
parabolic rule 603
partial derivatives, approximate calculation of 588
partial quotients 56
partitioned matrix (matrices) 256
  addition of 257
  conformal 257
  multiplication of 258
  subtraction of 257
partitioning, matrix inversion by 260
path 668
periodicity class 214
perpendicularity 345
Perron, O. 76, 677
Perron's theorem 390
point(s)
  interpolation (*see* interpolation points) 518
  mesh 518
  of n-dimensional space 336
polynomial(s)
  characteristic 376, 382
  Chebyshev 554
  computing values of 77
  left 241
  Legendre 611
  real roots of (number of) 173
  right 241
  Taylor 89
positive definite quadratic form 312
positive definite symmetric real matrix 388
positive definiteness, property of 343
postmultiplication 236
postmultiplier 269
power
  generalized 517
  relative error of 41
power series
  accelerating convergence of (by Euler-Abel method) 209
  method of 504
powers-of-ten notation 18, 24
premultiplication 236
premultiplier 269
principal element(s) 287
  method of 287
principal row 287
principle
  of the argument 166
  of equal effects 45
Pringsheim theorem 71
probabilities, transition 667
probability error estimate 52
probability integral 561
problem of interpolation (statement of) 518
procedure, direct and reverse 280
process
  Hero's 107
  Newton (*see* Newton process) 465

  root-squaring 182
  Seidel (*see* Seidel process) 155
product
  error of 37
  of matrices 231
  number of correct digits in a 39
  scalar (of vectors) 343
    properties of 343
projection transformation 369
propagation law of the ε-error 515
proper integral 633
proportional parts, method of 122
pseudorandom numbers 653
pseudorandom sequence 654
Purcell, method of 320

quadratic form 311
  matrix of 311
  negative definite 312
  positive definite 312
quadrature, mechanical 590
quadrature formulas 590
  accuracy of 618
  Chebyshev's 607ff
  of closed type 591
  Gauss' 611, 614
  Newton-Cotes 593
  Newton's 599
  of open type 591
quantity, random 651
quasidiagonal matrices 256
quotient(s)
  error of 41
  number of correct digits in a 41
  partial 56

random numbers 650
  generating 653
  tables of 653
random quantity 651
random sequence 651
  Gaussian 652
randomizing devices 653
rank of a matrix 248
rational fractions, computing values of 82
real matrices 389
reciprocals
  computing 104
  of square roots, computing 111
rectangular matrix 229
relations, biorthogonality 390, 393
relative error 19, 21
  of approximate number and number of correct digits 27ff
  of a power 41
  of a root 41
relaxation, method of 313ff.
remainder term 89
  of Bessel's interpolation formula 552
  Simpson's formula and its 596
  of Stirling's interpolation formula 552
  trapezoidal formula and its 595
Remez, E. Ya. 573, 677
residual 285
residual error 23
reverse procedure (*see* back substitution) 280
Richardson, method of 320
Richardson extrapolation 622ff
Rabinowitz, P. 674, 675
Robinson, G. 161, 272, 573, 678
Rolle's theorem 133
root(s)
  characteristic 375
  complex (*see* complex roots) 190
  of an equation 115
  existence of (of a system) 469
  finding (by inverse interpolation) 564
  -fold (*s*-fold) 178
  improving 284
  isolation of 115
  latent 375
  multiplicity of 162
  real (bounds of) 167
  real (of a polynomial) 173
  relative error of 41
  separated 180
  of a system of equations 274
root squaring 182
root-squaring process 182
rotation transformation 370
rounding of numbers 26ff
rounding errors 29
rounding-off rule 26
row vector 229
rule
  Cramer's 273
  even-digit 26
  parabolic 603
  rounding-off 26
  of signs, Descartes' 178
  Simpson's 597
  three eighths 599
  trapezoidal 601
Runge, C. 589, 677

Salekhov, G. 228, 677
Salvadori, M. G. 321, 648, 677
sampling error 666
scalar 229
scalar product(s)
  method of 436
  of vectors 343
    properties of 343
Scarborough, J. B. 54, 161, 321, 506, 573, 589, 648, 677
scheme
  Horner's 77, 78
    generalized 80-81
  of Khaletsky 295, 297
  of unique division 280
Schreier, O. 272, 677
scientific notation 18, 24
secular determinant(s) 376, 565
  expansion of 410ff
    comparison of different methods of 429
secular equation 376
Seidel method 309ff
Seidel process 155
  convergence of 327, 330, 333
  estimating the error of approximations in 330
    by l-norm 332
    by m-norm 330
separated roots 180
sequence
  convergent (of matrices) 249
  Gaussian random 652
  pseudorandom 654
  random 651
  Sturm 174
series
  convergence of, acceleration of 89
  Fourier trigonometric 213

  Maclaurin's 89
  matrix, sum of 251
  numerical 83
    convergent 83
  power (*see* method of power series) 504
  sum of 83
  Taylor's 89
sign 86
Shabat, B. V. 202, 675
Shapiro, G. M. 202 677
Shestakov, V. I. 114, 676
Shilov, G. E. A. 393, 677
Shreider, Yu. A. 393, 674, 677
Shura-Bura, M. R. 114, 676
sign(s)
  Descartes' rule of 178
  variations of (*see* variation of sign) 174
sign changes 174
significant digit 24
signum function 86
similar matrices 339, 380
  symbol for 339, 380
simple continued fraction 56
Simpson type cubature formula 644
Simpson's cubature formula 645
Simpson's formula and its remainder term 596
Simpson's general formula 603
Simpson's rule, formula of 597
singular linear transformation 374
singular matrix 236
singular transformation 375
singularities, isolation of by Kantorovich method 635
Smirnov, V. I. 114, 409, 677
Smolitsky, Kh. L. 321, 677
solution(s)
  approximate (of systems of nonlinear equations) 459
  of a difference equation 198
  fundamental system of 365
  graphical 119
solution set (of a system of equations) 274
solution space of a homogeneous system 364f
space
  basis of 340
  generated by vectors 342
  linear vector 336ff
  n-dimensional 336
    point of 336
    vector of 336
  n-dimensional real 344
  solutions 364
  spanned by vectors 342
spacing (*see* interval) 507
span (spanned by vectors) 342
Sperner, E. 272, 677
square matrix 229
square roots, computing 107
square-root method 293
standard continued fraction 56
steepest descent, method of 496, 499
  for a system of linear equations 501
Steffensen, J. F. 599, 648, 677
Stenin 465
Stirling, J. 205
Stirling's expansion 207
Stirling's interpolation formula 533
Sturm method 173
Sturm sequence 174
Sturm's theorem 174
submatrix (submatrices) 256
subspace
  dimensionality of 342
  linear (*see* linear subspace) 341
substitution
  back 280
  forward 280
subtraction, loss of accuracy in 35
sum(s)
  check 281
  error of 33
  of a series 83
Sylvester conditions 389
symmetric matrix (matrices) 235, 384
  properties of 384
symmetry, Hermitian 343
system(s)
  fundamental (of solutions) 365
  linear (*see* linear system)
  of linear algebraic equations, solution of by Monte Carlo method 666
  of linear equations
    solution of by orthogonalization methods 358
    solving 273ff
  of nonlinear equations, approximate solution of 459ff
  normal 311
  normal linear 312
  orthogonal 345

table(s)
  Barlow's 113
  of central differences 530
  for determining limiting relative error 30
  difference (*see* difference table) 510
  of divided differences 556
  double-entry 569
  of random numbers 653
    compilation of 653, 654, 656
tangents, method of 127
Taylor polynomial 89
Taylor's series 89
  term of κth component of continued fraction 55
  remainder 89
Ter-Mikaelyan, T.M. 114, 676
test, Cauchy's 83
theorem
  of Budan-Fourier 175, 176
  convergence 214
  Cayley-Hamilton 397
  fundamental (of algebra) 162
  Hua's 179
  Hurwitz 406
  Kantorovich 465
  Lagrange's 168
  Newton's 171
  Ostrowski's 159, 465
  Perron's 390
  of Pringsheim 71
  Rolle's 133
  Sturm's 174
  Weierstrass 389
theory
  of continued fractions 55ff
  of errors 42
    inverse problem of 44
      second 48

three-eighths rule 599
Tolstov, G. P. 114, 161, 228, 677
trace of a matrix 377
trajectory 668
transformation(s) 367
  of coordinates of a vector under changes in the basis 348
  Euler-Abel 210
  identical 375
  inverse 373
  Kummer 203, 204
  linear (of variables) 367
    degenerate 374
    operating with 371
    singular 374
  by a matrix, properties of 381
  nonsingular 374
  projection 369
  rotation 370
  singular 375
transformation matrix 368
transition probabilities 667
transition probability matrix 667-668
transpose of a matrix 234
  properties of 235
trapezoidal formula and its remainder term 595
trapezoidal rule 601
triangular matrices 265
trigonometric approximation 225
trigonometric functions 49, 98
  computing values of 98

undetermined coefficients, method of 411
unique division, scheme of 280
unit matrix 230

Vallée-Poussin, C. J. de la 228, 677
Vandermonde determinant 592, 614
variable interval 555
variations of sign 174
  lower number of 175
  upper number of 176
vector(s)
  angle between two 345
  column 229
  coordinates of 336, 341
  length of 345
  linear dependence of 337
  linearly dependent 337
  linearly independent 337
  of n-dimensional space 336
  negative of 336
  orthogonal 345
  orthogonal systems of 345
  product of 337
  product of by a scalar 337
  row 229
  scalar product of 343
    properties of 343
  zero 336
Ventsel, D. A. 54, 161, 677
Ventsel, E. S. 54, 161, 506, 674, 677
very much less than ($\ll$) 22
volume, algebraic 656

Wayland, Harold 458, 678
Weierstrass theorem 389
Whittaker, E. T. 161, 272, 573, 678
Willers, F. A. 465

Young's modulus 44

Zaguskin, V. L. 202, 678
zero of a function 115
zero matrix 230
zero vector 336
Zhidkov, N. P. 321, 458, 648, 675

Mir Publishers would be grateful for your comments on the content, translation and design of this book.

We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics